Inaugural - Dissertation

submitted to the

Combined Faculty of Mathematics,
Engineering and Natural Sciences

of

Ruprecht–Karls–University,
Heidelberg

for the degree of

Doctor of Natural Sciences

put forward by

M.Sc. Titus Leistner

born in Schlema, Germany

Date of oral exam:

Deep Learning-Based Depth Estimation from Light Fields

supervised by

Prof. Dr. Carsten Rother

Abstract

Light fields have emerged as a highly accurate method for depth estimation, known for its precision and robustness against occlusions. After the decline of consumer-based light field cameras, new industrial and research applications have emerged with very different demands, including the usage of high-resolution wide-baseline camera arrays and the need for a reliable confidence measure. This thesis responds to these evolving requirements with two main contributions: First, the introduction of EPI-Shift, a deep learning-based framework for depth estimation from both, small- and wide-baseline light fields. EPI-Shift combines discrete disparity classification with continuous disparity-offset regression and performs well on wide-baseline light fields, even when trained solely on narrow-baseline data. The second contribution focuses on multimodal posterior regression in depth estimation, useful for dealing with reflective and semi-transparent surfaces and for uncertainty quantification. This thesis contributes three deep learning-based approaches for depth posterior regression: Unimodal Posterior Regression (UPR), EPI-Shift Ensemble (ESE), and Discrete Posterior Prediction (DPP). Each of these methods displays strengths and weaknesses for different applications, evaluated using a novel multimodal light field depth dataset. Even with the extended applicability to wide-baseline light fields and the enhanced posterior regression capabilities, the performance of the presented methods stays on par with other state-of-the art approaches, marking a significant step towards practicality for today's applications.

Zusammenfassung

Lichtfelder haben sich als eine äußerst präzise Methode zur Tiefenmessung etabliert. Sie sind bekannt für ihre Genauigkeit und Robustheit an Verdeckungen. Nachdem der Markt für Verbraucher-Lichtfeldkameras geschrumpft ist, sind neue industrielle und Forschungsanwendungen mit gänzlich neuen Anforderungen entstanden. Dazu gehören die Kompatibilität mit hochauflösenden Wide-Baseline-Kamera-Arrays und die Quantifizierung von Unsicherheiten. Diese Dissertation adressiert diese neuen Anforderungen mit zwei Hauptbeiträgen: Zum einen mit der Einführung von EPI-Shift, einem deep-learning-basierten Framework zur Tiefenschätzung, das sowohl für Lichtfelder mit kleiner als auch großer Baseline geeignet ist. EPI-Shift kombiniert diskrete Disparitätsklassifizierung mit kontinuierlicher Regression der Restdisparität und funktioniert dadurch präzise auf Wide-Baseline-Lichtfeldern, selbst wenn es ausschließlich mit Daten von Lichtfeldern mit kleiner Baseline trainiert wurde. Der zweite Beitrag fokussiert sich auf die multimodale Posterior-Regression bei der Tiefenschätzung, was für den Umgang mit reflektierenden und halbtransparenten Oberflächen, sowie zur Quantifizierung von Unsicherheiten von Nutzen ist. Es werden drei Deep-Learning Ansätze zur Tiefenposterior-Regression etabliert: Unimodal Posterior Regression (UPR), EPI-Shift Ensemble (ESE) und Discrete Posterior Prediction (DPP). Jede dieser Methoden zeigt Vor- und Nachteile in verschiedenen Anwendungsgebieten, welche anhand eines neuartigen multimodalen Lichtfelddatensatzes bewertet wurden. Trotz der erweiterten Anwendbarkeit auf Wide-Baseline-Lichtfelder und den verbesserten Fähigkeiten in der Posterior-Regression, bleibt die Genauigkeit der vorgestellten Methoden auf dem Niveau anderer führender Ansätze. Damit sind die diese Methoden ein wichtiger Schritt hin zur praktischen Nutzung in aktuellen Anwendungen.

Acknowledgements

I want to thank Prof. Dr. Carsten Rother for being my supervisor and letting me join his excellent team. The valuable discussions, the freedom to explore my ideas, and the guidance on writing papers he provided have been instrumental in my growth as a researcher. I extend my thanks to Prof. Dr. Bernd Jähne for accepting the role of my second examiner, and to Prof. Dr. Stefan Gumhold for his insightful discussions and review of my first publication.

Special appreciation goes to Dr. Hendrik Schilling for introducing me to the topic of light field depth estimation, and to Radek Mackowiak and Dr. Lynton Ardizzone for their ideas, fruitful discussions, and their support. I also want to thank Philip Grassal for the excellent collaboration on various projects.

Further, I want to thank my parents for their continuous support. Above all, I reserve my deepest gratitude for Sina, whose support has been the foundation of my achievements throughout the last years.

Contents

Acronyms

3DV	International Conference on 3D Vision	3
AR	Augmented Reality	18, 56
ASI	Angular Sampling Image	43
AuSE	Area under the Sparsification Error	82, 84, 93
BN	Batch Normalization	66, 68, 81, 91, 92
BNN	Bayesian Neural Network	58
BpCNet	BadPix Correction and post-refinement Network	54
CE	Cross Entropy	78, 79
CNN	Convolutional Neural Network	18, 27, 29, 30, 63–65, 69, 71
CPU	Central Processing Unit	85
CT	Census Transform	44
CUDA	Compute Unified Device Architecture	55
CVPR	Computer Vision and Pattern Recognition Conference	3
DfD	Depth from Defocus	18, 27, 28, 31, 46, 47
DFSI	Dual-Focus Stereo Imaging	48
DKM	Defocus Kernel Map	48
DPP	Discrete Posterior Prediction	iii, v, 2, 74, 81–88, 91–93, 95–98
DSLR	Digital Single-Lens Reflex	40, 54
EPI	Epipolar-Plane Image	19, 22–26, 28–32, 35–40, 42, 43, 45, 46, 49–52, 61–66, 69–71, 74, 75, 77, 78, 81, 84, 87–89
ESE	EPI-Shift Ensemble	iii, v, 2, 73, 74, 77–79, 81, 82, 84, 86–88, 91–93, 95–98
GMM	Gaussian Mixture Model	59
GPU	Graphics Processing Unit	57, 68, 85, 87

GRAD	Sum of Gradient Differences	41
IB	Information Bottleneck	59
IBR	Image-Based Rendering	56, 57
INN	Invertible Neural Network	59, 60
KLD	Kullback-Leibler Divergence	76, 82–84
LCD	Liquid-Crystal Display	15
LIDAR	Light Detection and Ranging	17, 31, 61
MCMC	Markov Chain Monte Carlo	58
ML	Maximum Likelihood	74, 76
MPI	Multi-Plane Image	37, 57
MRF	Markov Random Field	46, 47
MSE	Mean Squared Error	69, 82–85, 93
MVS	Multi-View Stereo	19, 31
NeRF	Neural Radiance Field	6, 57
NP	Nondeterministic Polynomial Time	41
PCA	Principal Component Analysis	41
PSF	Point Spread Function	46
ReLU	Rectified Linear Unit	51, 66, 67, 81, 91, 92
RPCA	Robust Principal Component Analysis	40
SA	Simulated Annealing	47
SAD	Sum of Absolute Differences	41, 44
SCam	Surface Camera	41, 43
SE	Sparsification Error	82, 84
SfM	Structure from Motion	18, 31
SIFT	Scale-Invariant Feature Transform	41
SLFC	Sparse Light Field Coding	84, 85
SPO	Spinning Parallelogram Operator	42, 43, 69
SPP	Spatial Pyramid Pooling	52

TGV Total Generalization Variation 43

ToF Time of Flight 17, 31

UPR Unimodal Posterior Regression . . iii, v, 2, 73, 74, 81–84, 88, 91–93, 95–98

VFX Visual Effects 19

VGA Video Graphics Array 54, 55

VR Virtual Reality 16, 19, 56, 73

ZNCC Zero-mean Normalized Cross Correlation 44

Symbols

b	baseline	21, 22
$C^{s}_{\text{true}}(x, y)$	ground truth classification output for light field shifted by s	67–69
$C^{s}(x, y)$	classification output for light field shifted by s	64, 65, 67, 68
d	disparity	21, 22, 65, 74–79, 82–84, 93–98
$D_{\text{int}}(x, y)$	discrete integer disparity map	65
$D_{\text{true}}(x, y)$	ground truth disparity map	67, 68
$D(x, y)$	disparity map	64, 65, 67
η	opacity	75, 76, 80, 83
f	focal length	22
f_w	neural network with weights w (probabilistic notation)	75–79
$\mathcal{L}$	loss function	67, 68, 76–79, 81–85, 89, 93
l	light field (probabilistic notation)	74–79, 83
λ	wavelength	6
$L^{s}_{u}(x, y, v)$	vertical Epipolar-Plane Image (EPI) stack, shifted by s pixels	65
$L^{s}_{v}(x, y, u)$	horizontal EPI stack, shifted by s pixels	65, 77
$L_{u,v}(x, y)$	view	8, 25
$L_{x,u}(y, v)$	vertical EPI	22, 24
$L(x', y', u', v')$	two-plane four-dimensional light field	6, 7
$L(x, y, u, v)$	image-view four-dimensional light field	7, 24
$L(X, Y, Z, \theta, \phi)$	five-dimensional light field	6
$L(X, Y, Z, \theta, \phi, \lambda, t)$	seven-dimensional light field (plenoptic function)	6
$L_{y,v}(x, u)$	horizontal EPI	22, 25

$(R, G, B)^\top$	Color	6, 7
$R^s(x, y)$	regression output for light field shifted by s	64, 65, 68
s	shift	65, 67, 68, 77
σ^2	uncertainty (variance)	81, 94–98
t	time	6
$(\theta, \phi)^\top$	orientation	6
$(u', v')^\top$	near-plane coordinates	7
$(u, v)^\top$	view coordinates	7, 8, 10, 12, 23, 65
w	neural network weights	75–77
$(x', y')^\top$	far-plane coordinates	7
$(x, y)^\top$	image coordinates	7, 8, 12, 21, 23, 65, 67, 68
$(X, Y, Z)^\top$	three-dimensional coordinates	6, 22
Z	depth	21, 22

1. Introduction

This chapter motivates and summarizes the contributions presented in this thesis.

1.1 Motivation

Depth from light fields stands as arguably the most accurate camera-based method for depth estimation [1], [2]. The high amount of redundancy captured in light fields not only leads to high depth accuracy, but also robustness at occlusions.

The first light field recording systems were typically gantries with a single camera or large camera arrays [3]–[6]. These wide-baseline systems are characterized by their exceptional accuracy. Later, following the introduction of integrated consumer plenoptic camera systems such as Lytro [7], the focus in the community shifted towards methods optimized for these devices [6], [8]–[12]. As a result, datasets and methods were predominantly tailored for narrow-baseline cameras. However, these cameras did not sustain long-term dominance in the market, mainly due to the low image resolution and high price [13]–[15]. In today's context, industrial applications of light field technology have gained much more significance [16]–[18]. Light field data also proves ideal for training deep neural networks for more cost-effective systems, such as stereo cameras, due to its high depth estimation accuracy [1], [19]. Despite this, most existing methods still prioritize the more readily available narrow-baseline training data [6], [10], which is modeled after plenoptic cameras, rendering these methods unsuitable for wide-baseline application fields.

Moreover, there is arguably a tendency for methods to become overly optimized for existing benchmarks. For instance, Shin *et al.* [20] and Tsai *et al.* [21] manually removed all specular areas from the training dataset, aiming to boost performance on the validation dataset that lacks any specular reflections. However, in practical and industrial applications, trustworthiness and reliable uncertainty quantification take precedence over raw depth accuracy, which might be narrowly overfitted to a specific benchmark. Most downstream tasks, like visual odometry, benefit more from accurate depth uncertainty than from perfectly smooth outputs in ambiguous areas [22]. Furthermore, current benchmarks do not address challenges such as semi-transparency and specular reflections, leaving these critical issues unresolved.

Based on these considerations, this thesis aims to adapt light field depth estimation to meet the demands of modern applications. We introduce EPI-Shift, a method that enables deep learning-based depth estimation to be applied to wide-baseline light field cameras, suitable for industrial applications or for gathering ground-truth depth data. Additionally, the thesis introduces and compares methods for depth posterior regression to enable uncertainty quantification for light field depth estimation and improve the performance on semi-transparent and reflective surfaces.

1.2 Contributions

The goal of this thesis is to enhance the practicality and applicability of light field depth estimation to today's applications.

In the first contribution, a novel deep learning-based framework is introduced, designed for both small- and wide-baseline light fields. We therefore contribute a novel deep learning-based framework we term *EPI-Shift*. *EPI-Shift* allows the network to handle a broad range of disparities efficiently while still operating within a small receptive field. The framework combines classification of discrete integer disparities with regression of continuous disparity-offsets. It is based on a U-Net component, which plays a crucial role in enhancing long-range smoothing and reducing artifacts. Our approach achieves performance comparable to existing deep learning-based state-of-the-art methods, while extending its applicability to wide-baseline inputs, despite being trained exclusively on narrow-baseline data. Its effectiveness is validated through experiments on both a publicly available synthetic small-baseline benchmark and large-baseline real-world recordings.

The second contribution of this thesis addresses depth posterior estimation from light fields. This is particularly relevant for uncertainty quantification and in scenarios involving reflective and semi-transparent surfaces, where other depth estimation methods are often inaccurate. Existing methods can not handle multiple depths, due to only modelling a single true depth per pixel, ignoring the likelihood of multiple objects at varied depths influencing the pixel's color. In response, we contribute a novel approach that outputs a posterior depth distribution instead of a single depth estimate. We introduce three novel deep learning-based depth estimation techniques, each capable of handling multiple depth modes. We term these techniques *Unimodal Posterior Regression (UPR)*, *EPI-Shift Ensemble (ESE)*, and *Discrete Posterior Prediction (DPP)*. Another significant contribution is the creation of the first-ever *multimodal light field depth dataset*. This dataset captures the depths of all objects that contribute to the color of each pixel in an image. An extensive evaluation of the predicted depth posterior distributions has led to insightful findings: The *UPR* method, while effective under the traditional unimodal depth assumption, becomes less accurate when this assumption is invalid. Contrastingly, the DPP method generally outshines both other methods for multiple depth modes, while ESE estimates accurate uncertainties for wide-baseline light fields.

1.3 Foundational Articles

This thesis is based on the following articles:

Learning to Think Outside the Box:
Wide-Baseline Light Field Depth Estimation with EPI-Shift
T. Leistner, H. Schilling, R. Mackowiak, S. Gumhold, C. Rother
International Conference on 3D Vision (3DV) 2019

Towards Multimodal Depth Estimation from Light Fields
T. Leistner, R. Mackowiak, L. Ardizzone, U. Köthe, C. Rother
Computer Vision and Pattern Recognition Conference (CVPR) 2022

During my PhD, I also contributed to this article not included in the thesis:

Neural Head Avatars from Monocular RGB Videos
P. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, J. Thies
Computer Vision and Pattern Recognition Conference (CVPR) 2022

1.4 Outline

The subsequent chapters introduce light field depth estimation, review relevant literature and detail our novel contributions to the field.

Chapter 2 — Background provides an in-depth overview of light field technology. The chapter starts with a discussion of the plenoptic function and various light field acquisition methods. It then briefly describes the most common methods for depth estimation and discusses sources of depth cues in light fields. The chapter concludes with a description of commonly used methods for depth estimation.

Chapter 3 — Related Work offers a literature review, targeted on the topics of this thesis. It starts with the various categories of light field depth estimation approaches and also includes surveys, benchmarks and datasets commonly used by the community. Subsequently, it provides a concise overview of light field rendering techniques, focusing on the foundational articles that introduced the key concepts. The final section of the chapter addresses deep learning-based methods for posterior regression which are particularly relevant for multimodal depth estimation, introduced in chapter 5.

Chapter 4 — EPI-Shift introduces a novel neural network framework for effective depth estimation from light field data. Compared to previous deep learning methods, *EPI-Shift* has the ability to handle wide-baseline light fields by virtually shifting the light field stack while maintaining a small receptive field and therefore low number of parameters. Our framework performs joint classification of discrete integer disparities and regression of continuous disparity-offsets. Experimental results show that EPI-Shift performs on par with existing learning-based and hand-crafted methods in both synthetic and real-world scenarios.

Chapter 5 — Multimodal Depth Estimation introduces different approaches for multimodal light field depth estimation, focusing on improving accuracy in semi-transparent and reflective areas. We shift from traditional singular depth estimates to outputting posterior depth distributions. These depth posteriors are useful to assess the trustworthiness of predictions and estimate multiple depth modes. Through comparative experiments, the chapter identifies the best methods for different applications, illustrating the strengths and weaknesses of each approach. Our findings demonstrate the potential of these deep learning-based methods to overcome long-standing challenges in the field.

Chapter 6 — Outlook and Conclusion discusses remaining challenges, suggests future research directions and applications for the light field depth estimation field and summarizes the thesis' key contributions.

2. Background

In this chapter, foundational concepts of light fields and depth estimation are explored. The chapter starts with a discussion of different mathematical models of vision, culminating in a definition of the four-dimensional plenoptic function, used throughout this thesis. Subsequently, various techniques for recording light fields, along with their advantages and disadvantages, are examined. Light field camera arrays, gantries, plenoptic cameras, programmable aperture cameras and synthetic light fields, are discussed.

The latter part of the chapter focuses on depth estimation, starting with an overview of general depth estimation techniques. Then, the discussion extends to different types of depth cues from light fields: Epipolar geometry is introduced for stereo camera setups and later extended to four-dimensional light field cameras. Additionally, plane sweep and defocus-induced depth cues are briefly discussed. The chapter concludes by describing common concepts across most depth estimation methods, spanning traditional optimization-based to deep learning approaches.

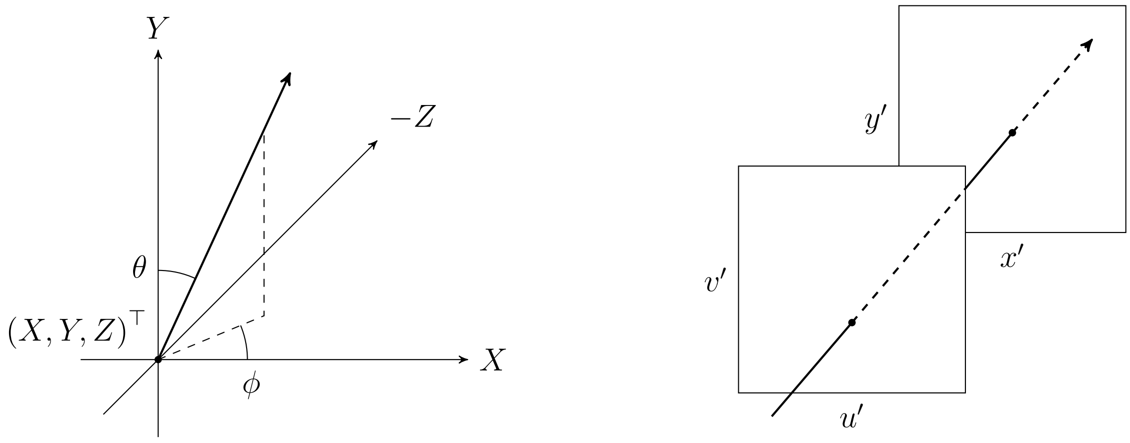
2.1 The Plenoptic Function

The exploration of human vision has captivated philosophers, researchers, and artists for centuries. Throughout history, the conceptualization of vision has evolved considerably.

The pinhole camera, with its origins in ancient China and Greece, exemplified fundamental optics principles through its elementary method of image projection. Arabic scholars, notably Alhazen, played an important role in the advancement of optics, significantly enhancing and preserving the optical theories established by the Greeks. In the 17th and 18th centuries, Kepler [24] and Newton [25] contributed groundbreaking work to the field. Their lectures and publications had a profound impact on the understanding of light.

In 1846, Faraday [26] introduced the concept of a *light field* in his lecture. This innovative concept envisioned a space filled with lines of force emanating from light sources. Faraday's theory sought to unify the phenomena of light with electromagnetism, proposing a novel linkage between light and the electromagnetic field. Although Faraday's specific concept of a light field did not entirely align with the emerging theories of quantum physics and photon theory in the early 20th century, it nonetheless found significant application in technical fields, such as computer graphics and computer vision.

The introduction of the *plenoptic function* by Adelson and Bergen [23] is a modern interpretation of Faraday's definition that mathematically models the intensity of light at all points in space and along all angles. This model is a simplification in



(a) 5D plenoptic function $L(X, Y, Z, \theta, \phi)$ (b) 4D plenoptic function $L(x', y', u', v')$

Figure 2.1: Two definitions of the plenoptic function. The five-dimensional definition (a) by Adelson and Bergen [23] models every location and orientation in space. The four-dimensional two-plane definition (b), used for most applications, is sufficient to model all rays in the empty volume between the planes. Reproduced from Levoy and Hanrahan [3].

the sense that light can only propagate along straight lines, an approximation of the actual complexities of the physical world. Nevertheless, this simplified model proves to be remarkably effective for a multitude of vision tasks, particularly in the context of light field photography. The authors introduce the concept of an imaginary eye or camera, positioned arbitrarily within space, capturing incoming light from all possible directions through a pinhole of infinitesimal size. The function that describes the radiance perceived by this eye is termed the *plenoptic function*

$$L(X, Y, Z, \theta, \phi, \lambda, t) . \tag{2.1}$$

For this definition, $(X, Y, Z)^\top$ denotes the three-dimensional position of the pinhole, $(\theta, \phi)^\top$ the orientation of the light ray, λ the wavelength of the light, and t the time.

In this thesis, a different definition of the plenoptic function, closer to practical computer vision applications, is used. Instead of utilizing the wavelength parameter λ, the codomain of L is defined as the red, green, and blue color components of light $(R, G, B)^\top$, which are commonly used in digital photography. Additionally, for the sake of simplicity, the time component is omitted, with the assumption of a singular, fixed moment in time. For applications like light field videos, the plenoptic function can be intuitively extended to include the time dimension. The domain of the newly defined plenoptic function comprises the pinhole's position in three-dimensional space $(X, Y, Z)^\top$ and the angular orientation of a light ray $(\theta, \phi)^\top$:

$$(R, G, B)^\top = L(X, Y, Z, \theta, \phi) . \tag{2.2}$$

This five-dimensional parameterization, as visualized in fig. 2.1a, has gained widespread acceptance in the field of computer vision. A notable application are Neural Radiance Fields (NeRFs), introduced by Mildenhall *et al.* [27], which involve training a neural network to represent the entire plenoptic function of a scene for the purpose of image-based rendering. Nonetheless, from a practical perspective, this model still has two inherent flaws:

The first problem is, that it does not consider constancy along rays. If we assume that there are no objects between two points, say A and B, the radiance of the ray pointing from A to B remains constant across the whole line segment $\overline{AB}$. Given that the majority of the real world is composed of empty space, modeling each and every point of the plenoptic function becomes a waste of memory and computational resources in most practical applications.

The second limitation is the impossibility of capturing a whole five-dimensional plenoptic function, as no existing device is capable of this. This becomes evident when considering Adelson's theoretical camera. In order to measure the plenoptic function at every location and direction in space, such camera would inevitably interfere with itself due to blocking the light from certain directions. Hence, the camera itself would obstruct the recording of the plenoptic function.

Therefore, it is only possible to measure the plenoptic function within a defined, limited volume of empty space by positioning the camera outside of this volume. This defined volume, typically shaped as a cube, cuboid, or pyramid frustum, accommodates the camera's field of view and is often defined as the space between two planes. Because this space is empty, the assumption of constancy along each ray between the two planes always holds true. Under these conditions, the domain of the plenoptic function can be reduced from five dimensions to four. Among the various parameterization methods available, the two-plane model is particularly prevalent. It assigns radiance values to each ray intersecting two planes, utilizing two-dimensional coordinates $(u', v')^\top$ on the first plane and $(x', y')^\top$ on the second plane. This two-plane plenoptic function can be written as

$$(R, G, B)^\top = L\left(x', y', u', v'\right) \tag{2.3}$$

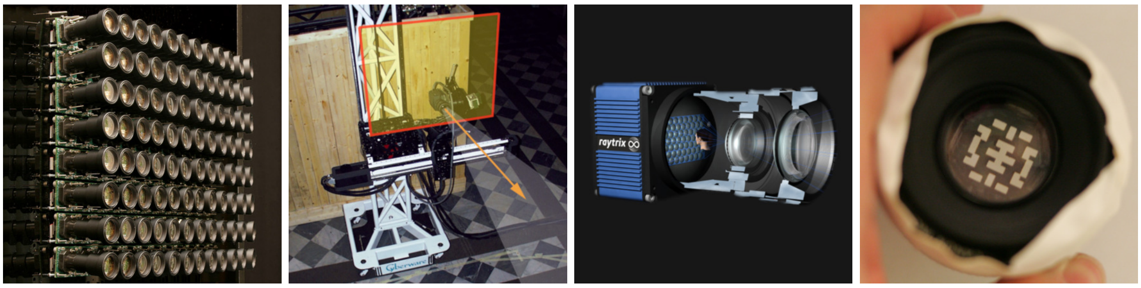
and is illustrated in fig. 2.1b.

This model conveniently already characterizes the majority of light field cameras used in today's applications. Imagine an array of identical pinhole cameras, all oriented in the same direction and positioned on a regular two-dimensional grid, as depicted in fig. 2.13. The pinholes of all cameras are located on the first plane, while their image planes, or sensors, are located on the second plane. Hence, each ray within the empty space can also be parameterized by the two-dimensional coordinate $(u', v')^\top$ of the camera's pinhole on the first plane and the coordinate $(x', y')^\top$ of the image pixel on the second plane.

However, out of practical considerations, it's beneficial to slightly adjust this parameterization. Rather than using the actual spatial position of each camera's pinhole, assigning integer grid positions $(u, v)^\top$ to each camera or *view* proves more useful. For instance, the central camera usually has the coordinates $(0, 0)^\top$, while a camera one row above and one column to the left would be positioned at $(-1, 1)^\top$. For the second plane, employing the pixel coordinate on each camera's sensor $(x, y)^\top$ is easier than using the global location $(u', v')^\top$, because it varies from camera to camera, given that the coordinate origin of each sensor is situated at a slightly different location. These considerations lead to the final plenoptic function definition

$$(R, G, B)^\top = L\left(x, y, u, v\right) \tag{2.4}$$

with $(u, v)^\top$ being the view (or camera) coordinate and $(x, y)^\top$ being the image (or pixel) coordinate. This definition will be used for the remainder of this thesis.



(a) Camera Array [28] (b) Camera Gantry [29] (c) Plenoptic Camera [30] (d) Coded Aperture [31]

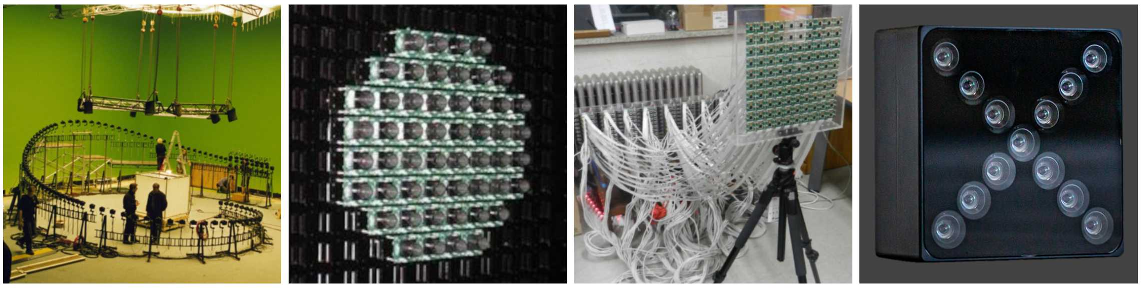
Figure 2.2: Different types of planar light field cameras. Note the difference in complexity, size and cost, making the plenoptic Raytrix camera (c) and coded aperture camera (d) more convenient but also reducing the resolution and overall image quality compared to camera arrays (a) and gantries (b).

2.2 Light Field Photography

This section introduces practical approaches to capturing light fields. There are two primary types of light field cameras: planar and spherical. Planar light field recording, historically introduced by Lippmann [32] in 1908 and later developed by Adelson and Wang [33] in 1992, captures light information on a two-dimensional plane and is the more common approach. In contrast, spherical light field recording encompasses capturing light from all directions around a point, similar to 360° photography, and is instrumental in creating immersive environments. Notable works in this domain include those by Ihm *et al.* [34], Shum and He [35], Overbeck *et al.* [36] and Broxton *et al.* [37]. Despite the vast potential of spherical light fields for immersive experiences, they remain a niche application, primarily due to the complexity and impracticality of the capture hardware. Planar light fields, in comparison, are widely adopted because they offer a balance between advanced imaging capabilities and ease of capture. Therefore, this section will focus on planar light field recording systems. It starts with a formal definition of a planar light field camera. Then practical implementations via camera arrays, gantries, plenoptic, and coded aperture cameras, as shown in fig. 2.2, are introduced. Lastly, synthetic light field rendering is briefly discussed.

2.2.1 Planar Light Field Camera

Before we delve into the practical aspects of light field recording techniques, it's crucial to first establish the concept of a planar light field camera, henceforth simply referred to as a *light field camera*. A light field camera can be fundamentally described as a collection of identical pinhole cameras, all oriented in the same direction and arranged on a regular two-dimensional grid. This is illustrated in fig. 2.13, using pyramids to visualize the pinhole cameras. The apex of each pyramid represents the pinhole, while the base, placed between the pinhole and the scene for clearer illustration, signifies the image plane or camera sensor. Each camera is identified by its grid position, or view coordinate, denoted as $(u, v)^\top$. When capturing a light field, the sensor of each individual camera records an image $L_{u,v}(x, y)$, indexed with $(x, y)^\top$. These images are often termed *sub-aperture images* in the context of plenoptic cameras, but for brevity this thesis will refer to them simply as *views*.



(a) *The Matrix* [38] (b) Wilburn *et al.* [39] (c) An *et al.* [40] (d) LumiScanX [18]

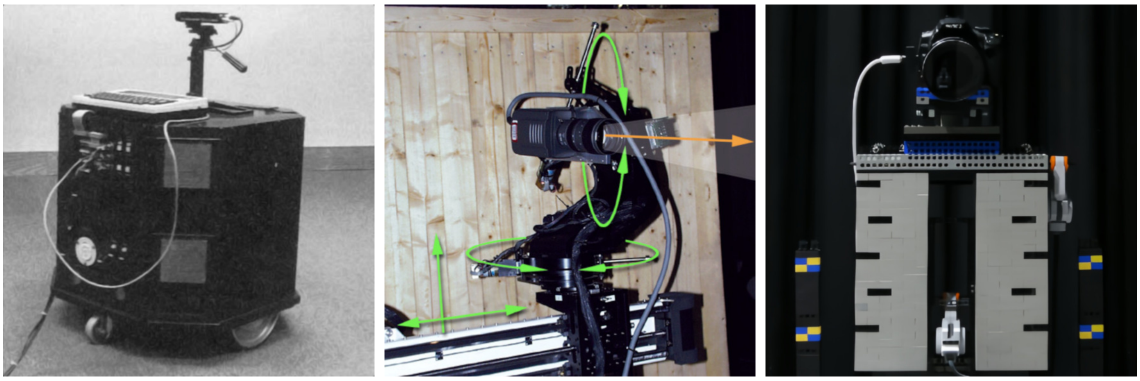
Figure 2.3: Evolution of light field camera arrays. Initially used for visual effects (a), diverse designs were proposed over the last 25 years (b, c). Today, integrated devices are used in industrial applications (d).

2.2.2 Camera Arrays

The array configuration represents the most direct implementation of our theoretical light field camera model. These setups typically involve cameras arranged in either a one or two-dimensional grid. Such an arrangement captures a three- or four-dimensional light field of a scene, with the cameras spaced at substantial distances from each other. A key advantage of camera arrays is the ability to simultaneously trigger all cameras, enabling the capture of both static images and videos of dynamic scenes. This does, however, require a system capable of processing and storing the high-bandwidth video stream.

The Matrix (1999) film's *bullet time* effect, a pioneering use of camera arrays, is a prime example of this principle [38]. The effect was produced using a circular array of cameras, shown in fig. 2.3a, allowing for a slow-motion effect with a dynamic, rotating perspective. This showcased the potential of camera arrays for creating visually impressive sequences. *The 3D Room*, developed by Kanade *et al.* [41] in 1998, utilized five light field camera arrays, one on each wall and one on the ceiling. Subsequent developments at Stanford University in the early 2000s saw the construction of various configurations of four-dimensional light field camera arrays, with one example illustrated in fig. 2.3b. The light field video array prototype presented by Wilburn *et al.* [42] showcased the feasibility of this model in practical applications, capable of compressing frames and recording from multiple cameras using a single PC. Wilburn *et al.* [39], [28] further explored the capabilities of these arrays, particularly in high-speed video recording (see fig. 2.3b), by alternating the triggering of cameras to achieve higher frame rates. More recently, An *et al.* [40] constructed a four-dimensional camera array, shown in fig. 2.3c, using accessible technology like Raspberry Pi computers [43], demonstrating the evolving accessibility of this technology.

Today, camera arrays have transitioned from research to industrial applications. Companies like *HD Vision Systems* have integrated light field arrays into devices such as their *LumiScanX*, shown in fig. 2.3d. It features 13 cameras in a cross-configuration and is used in applications such as industrial bin picking [16], [18] and quality assurance [17]. Moreover, light field arrays are also being utilized in the recording of training datasets for machine learning. One example is the company *rabbitAI* [19] providing a benchmark for depth prediction methods based on a dataset captured with a light field camera array [1].



(a) Bolles *et al.* [44] (b) Levoy and Shade [29] (c) Vaish and Adams [5]

Figure 2.4: Examples of light field gantry designs: A single-camera wheeled robotic platform for three-dimensional light field capture (a), a more advanced four-axis design (b) and an accessible gantry based on Lego Mindstorms [45](c).

While camera arrays are undeniably the most capable setups, enabling high resolution video recording of dynamic scenes, they also represent the most complex, bulky, and expensive light field camera configuration.

2.2.3 Camera Gantries

Light field gantries are a cost-effective alternative that requires only a single camera. The camera is mounted to a mechanized rig allowing for precise multi-axis movement. It is methodically moved to designated $(u, v)^\top$ locations and triggered each time to record one view. This method is laborious and restricts these setups to record static scenes only.

In 1987, Bolles *et al.* [44] developed an early example of a light field gantry (fig. 2.4a). Their system featured a single camera on a robotic wheeled platform, capable of capturing three-dimensional light fields of static scenes. Another light field gantry was constructed as part of the *Digital Michelangelo project* [29], [46]. It employed a computer-controlled mechanism for precise camera movements in four degrees of freedom: X and Y translations, nod, and shake, ensuring repeatable sub-millimeter accuracy. This setup was used to capture a light field of *Michelangelo's statue of Night* and is depicted in fig. 2.4b. The Lego Mindstorms Gantry, part of *the (new) Stanford Light Field Archive* by Vaish and Adams [5], depicted in fig. 2.4c, serves as an accessible example of light field gantries. It illustrates that capturing light fields doesn't require costly equipment to achieve accurate camera movement. The gantry, utilizing Lego Mindstorms motors [45], achieves relatively high accuracy and repeatability.

Despite their relative affordability and high resolution, the applications of light field gantries are limited. Their major drawback is the inability to capture videos or dynamic scenes, as the camera needs to move sequentially to multiple positions. Additionally, the moving parts might alter the scene lighting and cast shadows between views. Gantry systems also demand more complex mechanical controls and elaborate calibration, compared to camera arrays, which might be the main reason why camera gantries seem to be rarely used in industrial applications today.

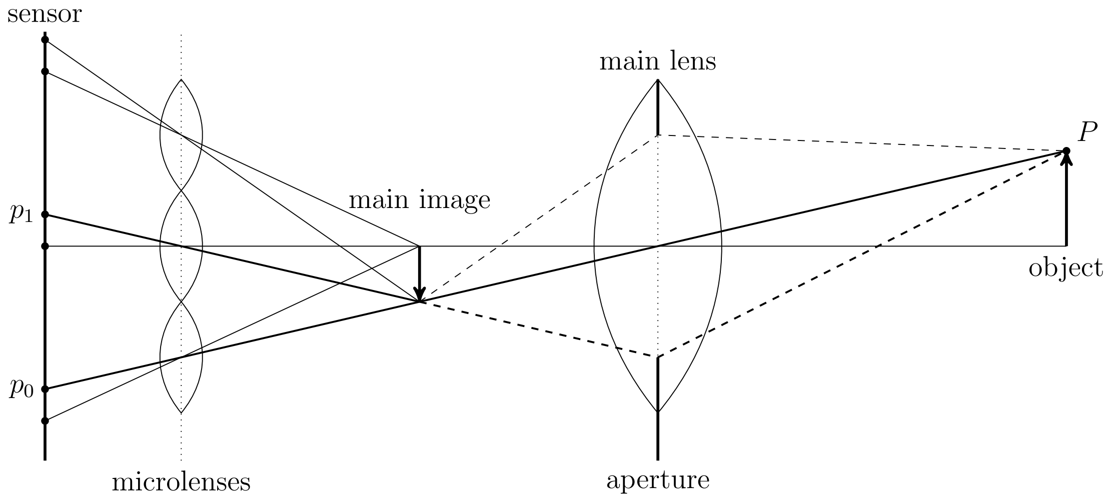


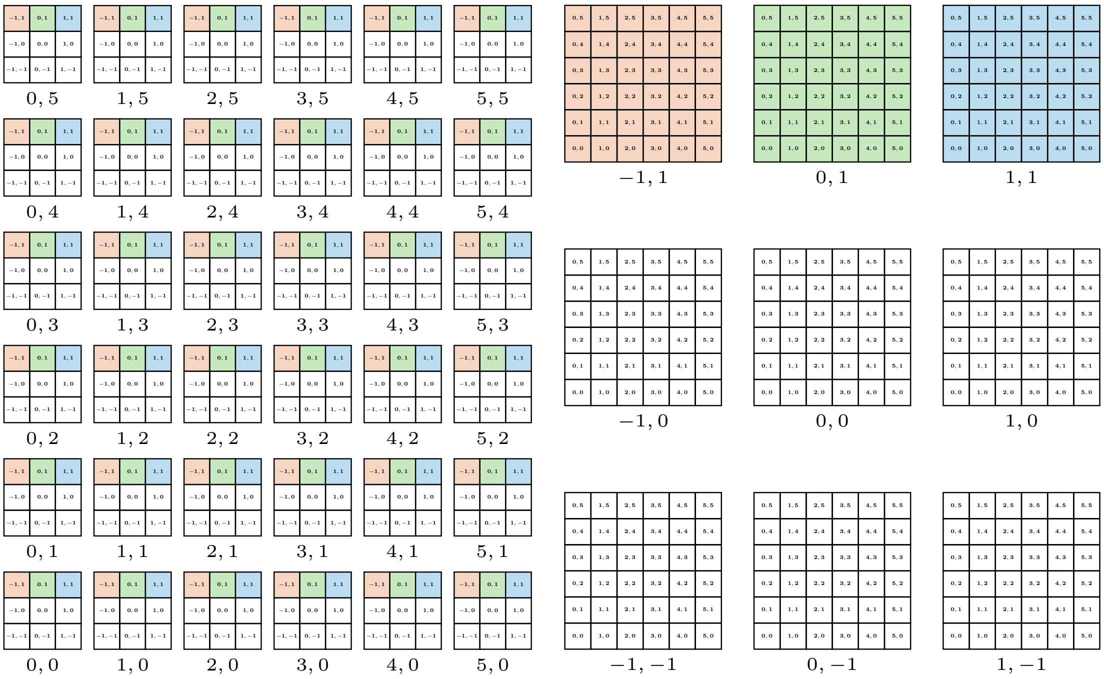
Figure 2.5: Simplified imaging process of a plenoptic camera. A single object point P is projected (thick rays) onto two distinct points p_0, p_1 in two separate microlens images. The specific position within these microlens images depends on the angle of incidence of the ray. Reproduced from Perwass and Wietzke [30].

2.2.4 Plenoptic Cameras

Plenoptic cameras were developed as a more accessible and compact alternative, extending their utility beyond specialized applications to consumer markets. Central to their design is a microlens array placed between the primary lens and the digital image sensor.

This microlens array enables the plenoptic camera to capture not merely a single *ray* of each point in the real world, but multiple rays originating from each point, along with their respective directions. Figure 2.5 illustrates a simplified plenoptic camera. Like a conventional digital camera, a real world object is focused onto a plane behind the main lens, termed *main image*. However, unlike standard cameras, the object's image isn't directly captured on an image sensor. Instead, a microlens array is positioned in front of the plenoptic camera's sensor. This configuration records the object's image at positions behind select or all microlenses, where a distinct *microlens image* materializes behind each one. The precise location of a light ray within the microlens image is dictated by the incidence angle of the ray. This effect enables the plenoptic camera to acquire four-dimensional light field information. For example, in fig. 2.5 both thick rays are reflected off the top of the real world object. These rays are refracted by the main lens towards different microlenses, one towards the center and one towards the bottom microlens. Note the projection of rays to different positions relative to the respective microlenses.

Unlike standard cameras, the image that is captured by the sensor of a plenoptic camera is made up of thousands of microlens images. To view the recorded views or to estimate depth similar to light fields obtained through arrays or gantries, it is necessary to first convert this collection of microlens images into views, which are often referred to as *sub-aperture images* within the context of plenoptic cameras. In practice, complex analysis and calibration algorithms are employed to identify each microlens image and adjust for various artifacts introduced by the microlens array. For an easier understanding, fig. 2.6 depicts this process in a simplified manner.

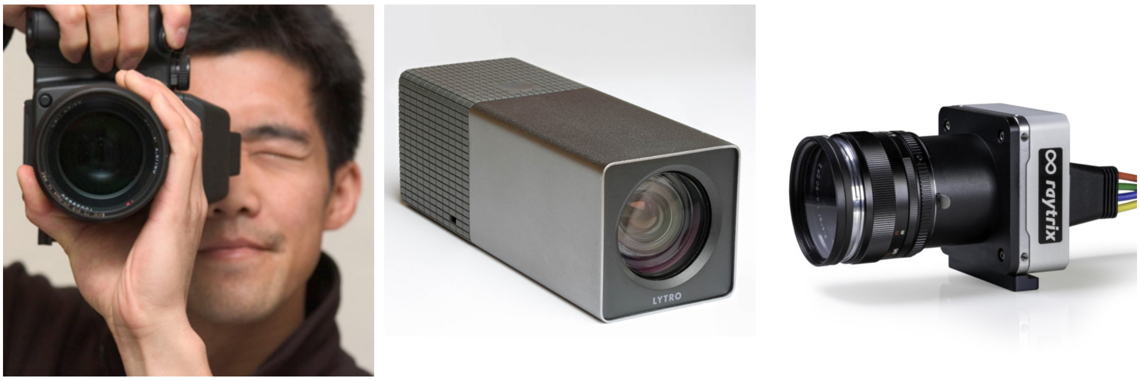


(a) Microlens images $(x, y)^\top$ (b) Views $(u, v)^\top$

Figure 2.6: Simplified illustration of micro lens images and views. A plenoptic camera records one microlens image per image coordinate $(x, y)^\top$ (a). The individual pixels can be re-arranged into views $(u, v)^\top$ (b). Reproduced from Hahne and Aggoun [47].

Figure 2.6a shows the image sensor located behind the microlens array, containing multiple microlens images arranged on a regular two-dimensional grid. Each microlens image represents one image coordinate $(x, y)^\top$ of the four-dimensional light field, whereas each pixel within a microlens image corresponds to one view $(u, v)^\top$. Put simply, the process of extracting views (see fig. 2.6b) from the microlens images is achieved by rearranging the pixels. One pixel from each microlens image is allocated to the first view, another to the second view, etc. The allocation is also illustrated using different colors for the first three views. For the real-world object point, depicted in fig. 2.5, the two highlighted rays are projected to different locations within the microlens images. This divergence is caused by the different angles of incidence into the camera and results in the allocation of these projections to different views. Consequently, similar to light field arrays, each view observes the scene from a slightly different perspective.

The concept of plenoptic imaging has roots tracing back to the early 20$^{\text{th}}$ century. It was first proposed by Lippmann [32], who introduced a precursor to the modern concept of capturing light fields. Lippmann's pioneering experiments yielded integral photographs that were relatively rudimentary in nature. These were produced using a plastic sheet embossed with a regular array of microlenses. Alternatively, he achieved similar results by embedding small glass beads, densely arranged in a random configuration, into the surface of the photographic emulsion. In the early 1990s, Adelson and Wang [33] introduced plenoptic cameras to the field of computer vision. They developed the first modern plenoptic camera, integrating the concept of a microlens array to capture light field information.



(a) Ng *et al.* [7]

(b) Lytro [7]

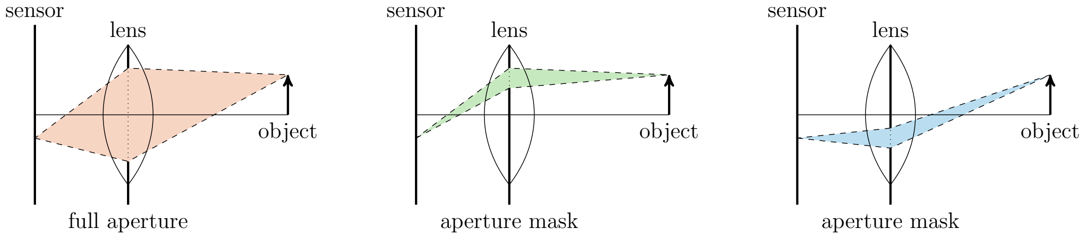
(c) Raytrix R65 [30]

Figure 2.7: Examples of plenoptic cameras. Ng *et al.* [7] presented an initial prototype (a) in 2005. He later founded Lytro, famous for its consumer light field camera (b). Raytrix (c) focuses mostly on industrial applications.

Lytro, founded by Ren Ng [7] in 2006, was a key player in bringing plenoptic technology to the consumer market. Lytro's first consumer camera, released in 2011, featured a novel design and allowed users to refocus images after capture, a feature that was groundbreaking at the time. After the release of their initial consumer product, Lytro went on to develop more advanced models, including the Lytro Illum in 2014, which targeted professional photographers with higher image quality and more powerful software for manipulating light field data. Despite their technological innovations, Lytro struggled commercially, primarily due to the limited resolution of images taken with its cameras. The technology's inherent division of sensor resolution to capture multiple light angles led to a decrease in overall image quality, a factor that significantly impacted consumer satisfaction. Additionally, the high cost of Lytro's cameras, when compared to conventional digital cameras offering higher resolutions, further hindered their widespread adoption [13]. The company first shifted its focus to virtual reality and video applications before ceasing operations in 2018 [14], [15].

Raytrix [30], founded in 2008, took a different path, focusing on high-resolution plenoptic cameras for industrial and scientific applications. Raytrix's cameras were designed to address some of the limitations of early plenoptic cameras, such as low resolution. They achieved this by using a complex microlens array design that allowed for higher resolution and greater depth accuracy. Raytrix's cameras found applications in various fields including biomedical imaging, face recognition and quality inspection [48].

In comparison to traditional camera arrays, plenoptic cameras offer the advantage of compactness and cost-effectiveness. This is because the microlens array replaces the need for multiple cameras to capture different perspectives of the scene. However, plenoptic cameras come with their own set of drawbacks. One of the main disadvantages is the lower spatial resolution, as the sensor's resolution is divided among capturing different views. Additionally, the low baseline of plenoptic cameras results in less accurate depth measurements, a limitation for applications requiring high-precision depth information.



(a) Conventional camera (b) Aperture mask 1 (c) Aperture mask 2

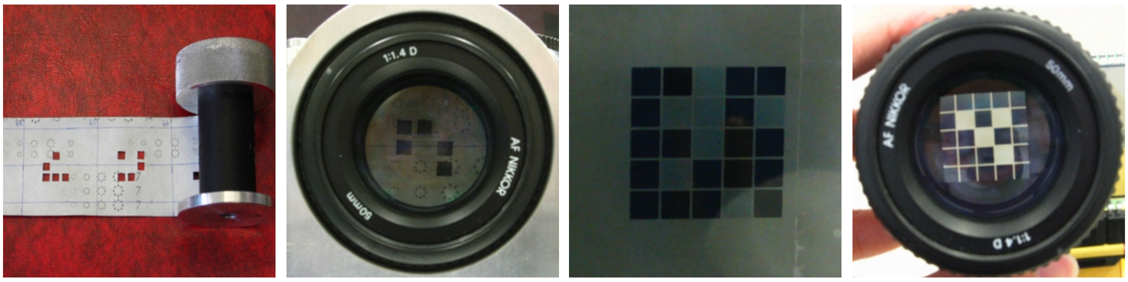
Figure 2.8: Simplified illustration of coded aperture light field photography. In conventional cameras, all rays emanating from an in-focus point are refracted to a single point on the sensor (a). Programmable aperture cameras employ aperture masks to restrict the angles of incoming light (b, c). Through the sequential capturing with varied aperture masks, these cameras are able to record a four-dimensional light field of the scene from different angles. Reproduced from Liang *et al.* [49].

2.2.5 Coded Aperture Cameras

To address the challenge of balancing spatial resolution with angular resolution, technologies such as *coded aperture* and *programmable aperture* cameras have been developed. Unlike plenoptic cameras, which rely on a microlens array, these cameras control the incoming light by incorporating various patterns within the camera's aperture.

The fundamental principle is depicted in Figure 2.8. Traditional cameras typically feature an aperture that is relatively symmetrical around the lens's center, allowing light to enter the camera uniformly from a broad range of angles. When an object in the real world is in the camera's focus, all rays emanating from this point and entering the camera are refracted by the lens to a single point on the image sensor, as illustrated in fig. 2.8a. Consequently, all angular information about these light rays is lost in conventional two-dimensional photography.

Programmable aperture photography, however, employs masks to cover parts of the aperture, only permitting light from specific angles to enter the camera. For ease of understanding, consider the aperture mask to have a single small hole, as depicted in fig. 2.8b and fig. 2.8c. This restriction narrows the range of incident angles significantly. Capturing an image with this hypothetical mask records the scene from that particular angle, akin to capturing a single view in a light field array or a view in plenoptic photography. By taking multiple photographs with different masks sequentially, one can record a complete four-dimensional light field. Nevertheless, the drawback of using a mask with a single hole is the significant reduction in light entering the camera, leading to noise in the resultant light fields. To overcome this, aperture patterns featuring multiple holes are utilized. Since the light hitting a specific pixel on the image sensor is simply the sum of light from all angles, the light from a particular angle can be computed from the images captured with various systematically selected aperture patterns. Figure 2.9 shows examples of such patterns. The alteration of patterns between shots can be achieved through methods such as the rapid scrolling of an opaque paper scroll (see fig. 2.9a) with precise electric motor control. A more compact alternative involves using a Liquid-Crystal Display (LCD) as the aperture mask (see fig. 2.9b) and refreshing the display between shots.



(a) Pattern scroll made from paper

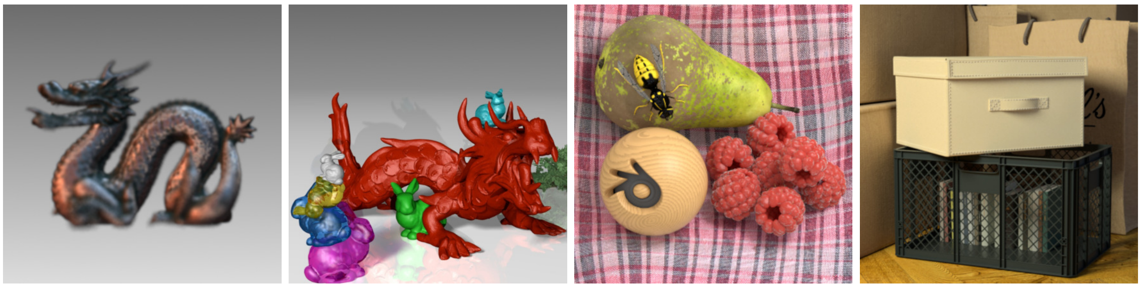
(b) LCD pattern

Figure 2.9: Different types of programmable aperture cameras, built by Liang *et al.* [49]. Changing the aperture pattern can be achieved by scrolling a slit of paper (a), or refreshing an LCD (b).

The concept of employing a coded aperture pattern for imaging purposes was initially proposed by Dicke [50] and Ables [51] in 1968, with further developments made by Fenimore and Cannon [52] in 1978. The technique has predominantly been utilized for X-ray and gamma-ray imaging to enhance light collection, as conventional glass lenses are ineffective at these wavelengths.

It was not until the 2000s that the method was extended to conventional photography, for refocusing, depth estimation, and light field recording. The first adaptation to this area was made by Levin *et al.* [31], who introduced a technique for refocusing and depth estimation using a single static coded aperture pattern. This work primarily focused on the design of an optimal aperture pattern to deduce the depth of real-world objects by analyzing the structure of blur in captured images. Their method is based on the assumption that the blur of a point on the image plane can be computed using a convolutional kernel, the size of which is dependent on the object's distance from the camera. Depth estimation is achieved by applying varying scales of the inverse convolution kernel locally and comparing the refocused image patch to a statistical model of real-world images. This process recovers the most probable depth for refocusing each image patch and, thus, also enables the reconstruction of a fully sharp image. However, because it only uses one fixed aperture pattern, the method cannot record full four-dimensional light fields. For the first time, full light field recording has been achieved by Liang *et al.* [49] by capturing a series of images with varying aperture patterns, as illustrated in fig. 2.9. This approach, akin to the capabilities of plenoptic cameras, can be used for depth estimation or refocusing with unknown depth.

Coded aperture cameras are considered superior to plenoptic cameras with respect to their ability to record at the sensors full pixel count, eliminating the compromise between angular and spatial resolution. However, the technique still faces a tradeoff between the quality of the light field and the number of images captured with different patterns. The feasible number of images is constrained by the refresh rate of the pattern and the potential for motion blur in dynamic scenes. The contrast of LCDs presents another limitation, as noted by Liu *et al.* [53]. Recent advancements have aimed to address these challenges by developing methods to reconstruct higher quality light fields from a reduced number of images, as demonstrated by Inagaki *et al.* [54], and proposing new hardware configurations, like Liu *et al.* [53]. Despite these innovations, coded aperture photography has yet to achieve widespread adoption in consumer or industrial cameras.



(a) Levoy and Hanrahan [3]

(b) Wetzstein [55]

(c) Wanner *et al.* [6]

(d) Honauer *et al.* [10]

Figure 2.10: Examples for synthetic light field datasets. Many datasets have been published, focusing on different aspects of light field technology: rendering (a), compression (b), and depth estimation (c, d).

2.2.6 Synthetic Light Fields

Synthetic rendering of computer-generated scenes offers an alternative approach to recording light field data. This approach improves real-time rendering and enables the creation of deep learning training data for practical light field applications.

In the field of computer graphics, synthetic light fields enable the pre-rendering of scenes. This pre-rendering permits the application of high-quality rendering techniques and visual effects that exceed the capabilities of real-time processing. The rendered light field data can subsequently be refocused and viewed from new perspectives efficiently in real-time, without sacrificing image quality and regardless of the scene's complexity. This feature is particularly advantageous for Virtual Reality (VR) applications, where high-quality, real-time rendering is essential. The concept of re-rendering synthetic light fields was initially proposed by Levoy and Hanrahan [3] in 1996. They provided the first synthetic light field dataset, consisting of four scenes rendered with different parameters. One of these scenes is depicted in fig. 2.10a.

Furthermore, synthetic light field data has been utilized for training and validating methods related to other light field technologies. For instance, Wetzstein [55] (see fig. 2.10b) introduced a synthetic dataset in his exploration of light field compression techniques. The significance of synthetic data increased, especially for the supervised training of neural network models. Especially the training of light field depth estimation methods depends hugely on synthetic data. Given the absence of scanning technologies capable of capturing depth maps of real-world scenes with comparable or superior density and accuracy, particularly for large scale or outdoor scenes, synthetic data has become indispensable. As a result, research focused on neural network based depth estimation from light fields today predominantly depends on synthetic training data. Key examples of depth validation and training datasets include those provided by Wanner *et al.* [6] and Honauer *et al.* [10] (see fig. 2.10c and fig. 2.10d). The dependency on synthetic data introduces a challenge in bridging the gap between synthetic training scenes and real-world recordings, with the goal of ensuring the accuracy and reliability of algorithms when applied to real images. Accordingly, more recent datasets focused on improving the photorealism and diversity of the generated data. Two notable examples are the datasets contributed by Li *et al.* [56] and Sheng *et al.* [12]. In conclusion, synthetic light field rendering plays a crucial role in the light field research community.

2.3 Overview of Depth Estimation Methods

This section offers a concise overview of various depth estimation methods used in computer vision. Depth estimation techniques are broadly divided into two main categories: Active methods, which involve the use of an artificial light source to illuminate the scene, and passive methods, which depend entirely on the light available in the scene.

2.3.1 Active Methods

Structured Light depth estimation is an active stereo vision technique. It operates by projecting a known pattern of light, often including lines, grids, or coded patterns, onto the scene. When this pattern interacts with the surface of the object, it deforms according to the object's geometry. A camera, positioned at a known pose relative to the light source, captures the deformed pattern. By analyzing the deviations from the known pattern, structured light methods can compute the depth information for each point on the object's surface, using triangulation principles. Posdamer and Altschuler [57] pioneered this technique, exploring how projecting patterns of light in specific ways can help measure an object's surface details. The approach is mostly used in applications requiring detailed surface topographies, such as quality control in manufacturing.

Time of Flight (ToF) sensors operate by sending out a, typically infrared, light signal towards objects and measuring the duration needed for the light to return to the sensor. The time it takes for this light to make the round trip is directly linked to the distance between the sensor and the objects it detects. Using this principle, ToF cameras are able to generate a depth map of the scene, calculating the distances to various points within their view. Nitzan *et al.* [58] contributed the earliest description of a laser-based ToF camera. Thanks to its minimal need for computational resources, ToF technology is especially useful in real-time applications.

Light Detection and Ranging (LIDAR) systems work by sending out laser beams towards a target and then timing how long it takes for each beam to bounce back after reflected by the surface. This method calculates the distance to an object using the known speed of light, factoring in the delay time between when the laser is sent out and when it returns. By moving the laser across an area and recording distances from many points, LIDAR creates a detailed three-dimensional map of the surrounding space. The first version of this technology was developed in 1961 by the *Hughes Aircraft Company*, as noted by Smith [59]. Today, LIDAR has a wide range of uses, including robotics and autonomous vehicles.

2.3.2 Passive Methods

Monocular Depth Estimation employs visual signals such as object size, texture gradients, occlusion, and perspective to deduce the distance of objects from the observer, capitalizing on the observation that objects appear smaller or more blurred are often further away. Convolutional Neural Networks (CNNs) are pivotal in this field, as they autonomously extract and learn features from images that signify depth. Saxena *et al.* [60] contributed the first method for monocular depth estimation and Ming *et al.* [61] provided a comprehensive review of more recent state-of-the-art methods. Given its cost-effectiveness, only necessitating a single camera, monocular depth estimation finds extensive application in autonomous vehicles and Augmented Reality (AR) applications.

Structure from Motion (SfM) operates by examining the relative movement of a camera navigating through an environment or through analyzing multiple images captured from different angles. It hinges on feature tracking, identifying and following distinct elements within the images, such as corners, edges, or keypoints, across a series of images. By observing the movement of these elements across frames, SfM can deduce the camera's pose at each point of image capture. Leveraging these determined poses, SfM triangulates the three-dimensional coordinates of the observed features, generating either a sparse or dense three-dimensional point cloud that models the scene's structure, where each point mirrors a real-world feature's position. The method was originally introduced by Ullman [62]. Owing to its proficiency in capturing large-scale environments, SfM is predominantly employed in fields like architectural and geographic scanning, as well as in robotics.

Depth from Defocus (DfD) determines the depth of objects in a scene by examining the blur or defocus observed in an image. This method leverages the correlation between blur intensity in an image and the object's distance from the camera, because varying distances lead to different levels of image blur. DfD usually requires taking several shots of the same scene with different focus adjustments, generating a series of images each exhibiting a distinct degree of defocus. Through analysis of these varying blur patterns and with knowledge of camera specifications, DfD algorithms are capable of estimating object depths, culminating in the creation of a depth map. The foundational work by Pentland [63] introduced the idea of utilizing focus variations to gauge scene depth, demonstrating that defocus signals can be methodically evaluated to deduce distances. DfD finds its utility in fields like robotics, photography, and computational photography, offering advancements in depth perception and facilitating effects such as artificial background blur in photos.

Stereo Depth Estimation leverages a pair of stereo images to derive three-dimensional depth information of a scene, mimicking human binocular vision with images from two cameras positioned slightly apart. The core concept of stereo depth estimation is triangulation, which uses the difference or disparity between matching points across the stereo images to calculate depth. Stereo cameras simultaneously capture two images from different angles, each providing visual information crucial for matching corresponding pixels between the left and right images. Depth calculation involves identifying similar feature points in both images through comparison of pixel intensity, patterns, or keypoints. Following the identification of matching points, a disparity map is created, indicating the pixel shift for each point between the two images, which directly relates to depth. Marr *et al.* [64] introduced one of the first computational models explaining depth perception from image disparities by human vision. A catalogue of state-of-the-art stereo methods is also provided by the ETH3D benchmark [65]. Stereo depth estimation's precision in generating three-dimensional models and depth data finds extensive use in robotics, autonomous driving, scene reconstruction, and object tracking.

Multi-View Stereo (MVS) extends stereo depth estimation to multi-camera systems. It hinges on triangulation, where the position of a point in space is determined by intersecting lines of sight from multiple cameras observing the same point. In MVS, a group of calibrated cameras captures overlapping images of the scene, enabling the identification of corresponding points or features in each image. These correspondences establish rays originating from the camera centers and passing through corresponding image points, forming a visual ray. By intersecting these visual rays in space, MVS algorithms calculate the location of observed scene points, forming a dense three-dimensional point cloud. Various MVS algorithms exist, including graph-based techniques, depth map fusion, and energy minimization approaches, each with its own strengths and trade-offs regarding accuracy and computational complexity. Okutomi and Kanade [66] first introduced the principles of using multiple images from distinct viewpoints for three-dimensional structure reconstruction. The ETH3D benchmark [65] also provides a catalog of state-of-the-art multi-view stereo methods. MVS finds applications in object scanning for Visual Effects (VFX), VR and more.

Light Field Depth Estimation, similar to MVS, calculates depth information from multiple images. However, there are certain distinctions between the two methods: Light field depth estimation uses either a single camera or a multi-camera arrangement to capture both spatial and angular characteristics of incoming light rays. Some light field cameras, like plenoptic and programmable aperture cameras are able to capture both spatial and angular light ray information and therefore estimate depth with just a single camera setup. In contrast, MVS relies on conventional cameras to capture multiple two-dimensional images of a scene from different viewpoints. Common techniques for light field depth estimation analyze line patterns in Epipolar-Plane Images (EPIs) to recover local disparities. In contrast, MVS employs multiple viewpoints and feature correspondences for depth estimation. Light field depth estimation typically yields a dense depth map for one or all views, while MVS produces either a dense or semi-dense point cloud representing the scene's geometry. MVS often demands significant computational resources and is computationally intensive, especially for dense reconstruction. Conversely, light field depth estimation is computationally lighter, making it suitable for real-time or resource-constrained applications. The concept of light field depth estimation was initially introduced by Bolles *et al.* [44], who recorded three-dimensional light fields using a mobile robot-mounted camera and conducted an early analysis of EPIs. Recent state-of-the-art methods can be found in the HCI 4D Light Field Benchmark [10]. Due to the compact nature of some light field camera models, this approach is frequently used in photography. Its computational efficiency and real-time capabilities also make it applicable in industrial and medical contexts. The precision of depth estimation from light fields also makes it a valuable method for generating training data for supervised training of stereo and monocular depth estimation algorithms.

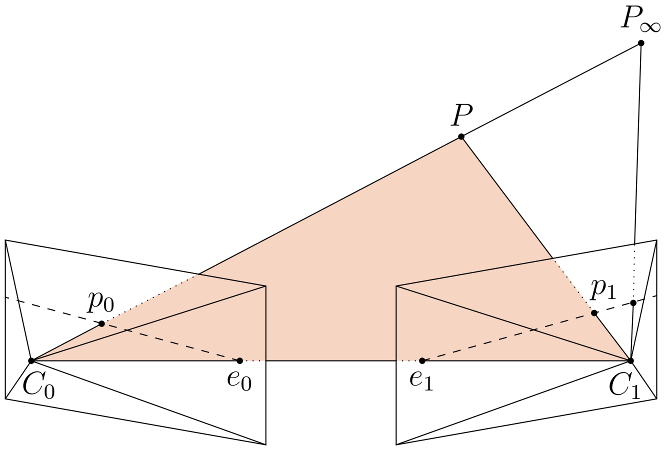


Figure 2.11: Visualization of epipolar geometry. Stereo depth estimation searches for the correspondence of a point p_0, captured by the first camera at C_0, in the image captured by the second camera at C_1. The correspondence p_1 always lies on the epipolar line between the epipole e_1 and the projection of p_0 at infinite depth P_∞. The orange plane is called *epipolar plane*. Reproduced from Szeliski [67].

2.4 Sources of Depth Cues

This section discusses the different sources of depth cues found in light fields, including epipolar geometry, plane sweep and defocus cues.

2.4.1 Epipolar Geometry

The most commonly adopted methods for depth estimation utilize triangulation, based on the principles of epipolar geometry. For clarity and simplicity, this concept is initially described using the example of a stereo camera setup, which can be viewed as a specific instance of a light field camera array, featuring just two cameras. The objective of stereo depth estimation is to recover the depth, or the distance between the camera pinhole C_0 and a three-dimensional point P, given the projection p_0 of this point in an image. When the projection p_1 of the same point P, recorded by the second camera at C_1 is available, depth can be determined through triangulation. It is assumed that the cameras have undergone calibration, making all essential intrinsic and extrinsic parameters accessible. The point p_1 is also termed the *correspondence* to p_0. The primary challenge in stereo depth estimation is the identification of this correspondence. In order to eliminate ambiguities and reduce computational cost, it's essential to constrain the search space in the second image as much as possible. Fortunately, potential correspondences are confined to a singular line, referred to as the *epipolar line*. This concept is illustrated in fig. 2.11. The epipolar line corresponding to a specific point p_0 is bounded by the epipole e_1, which is the projection of the first camera's pinhole C_0 onto the second camera's image plane. At the other end, it is bounded by the projection of p_0 at an infinite depth P_∞, depicted closer for illustrative clarity. The plane defined by C_0PC_1 (shown in orange) is termed the *epipolar plane*. The intersection of this plane with the two image planes yields a pair of corresponding epipolar lines. This information already significantly reduces the search space for correspondences.

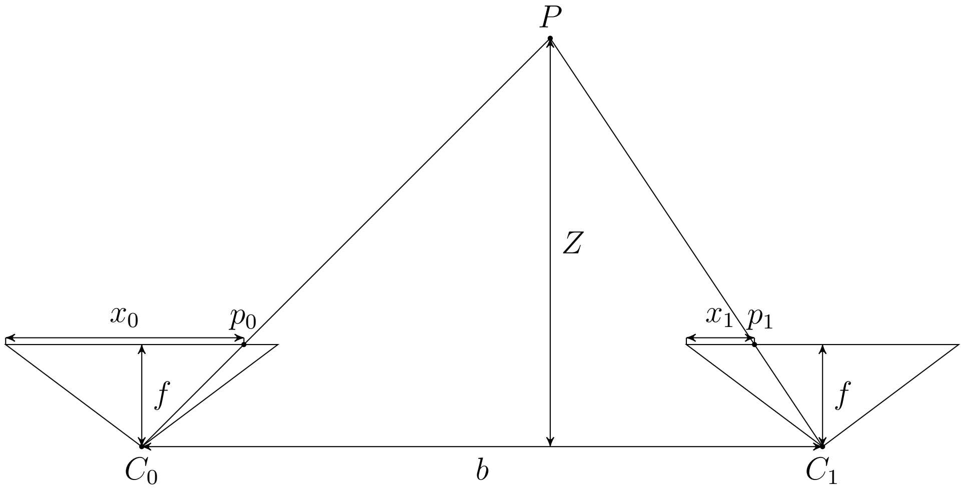


Figure 2.12: Standard rectified stereo geometry. Both cameras have the exact same intrinsics and are aligned perfectly parallel, offset by the baseline b. The depth Z of a point P can be computed using triangulation based on the disparity $d = x_0 - x_1$.

Rectification

For the majority of correspondence search algorithms, it is also beneficial to consider epipolar lines as coinciding with the horizontal pixel rows or vertical pixel columns within the captured images. Figure 2.12 depicts a bird's-eye view of a stereo camera configuration featuring horizontal epipolar lines. This implies that for any specific image coordinate $p_0 = (x, y)^\top$, its corresponding point p_1 is always located at the identical y coordinate, which significantly reduces the complexity of correspondence search algorithms. This configuration requires that two identical cameras are aligned perfectly parallel, which is almost unachievable in real-world camera assemblies.

To address this discrepancy, the recorded images are warped or *rectified*, based on the camera intrinsics and extrinsics, to ensure that they meet these prerequisites. A straightforward rectification technique was proposed by Fusiello *et al.* [68], involving three steps: Firstly, adjust the camera orientation to be orthogonal to the line connecting the camera centers. Secondly, establish the optimal rotation around the optical axes to ensure that the up-axis (the y-axis in image coordinates) is perpendicular to the line between camera centers, thereby guaranteeing horizontal alignment of corresponding epipolar lines. Lastly, adjust the image scale to compensate for variations in the focal lengths of the cameras, a frequent occurrence in real-world camera setups. Szeliski [67] describes the outcome as *standard rectified geometry,* the foundation for the majority of stereo depth estimation algorithms.

As illustrated in fig. 2.12, both *virtual* cameras are now perfectly parallel and have the same focal length, but with a horizontal offset, known as their baseline b. Epipolar lines align with the horizontal pixel rows, as defined by y, and the horizontal position x along these lines varies with X and the depth Z of a point in world coordinates. Assuming the projection p_0 of P and its correspondence p_1 have already been identified, the x-offset between these two projections in the images, is denoted as disparity $d = x_0 - x_1$. The depth Z bears an inverse relationship to the disparity. Given known camera parameters f and b, the equation relating depth Z

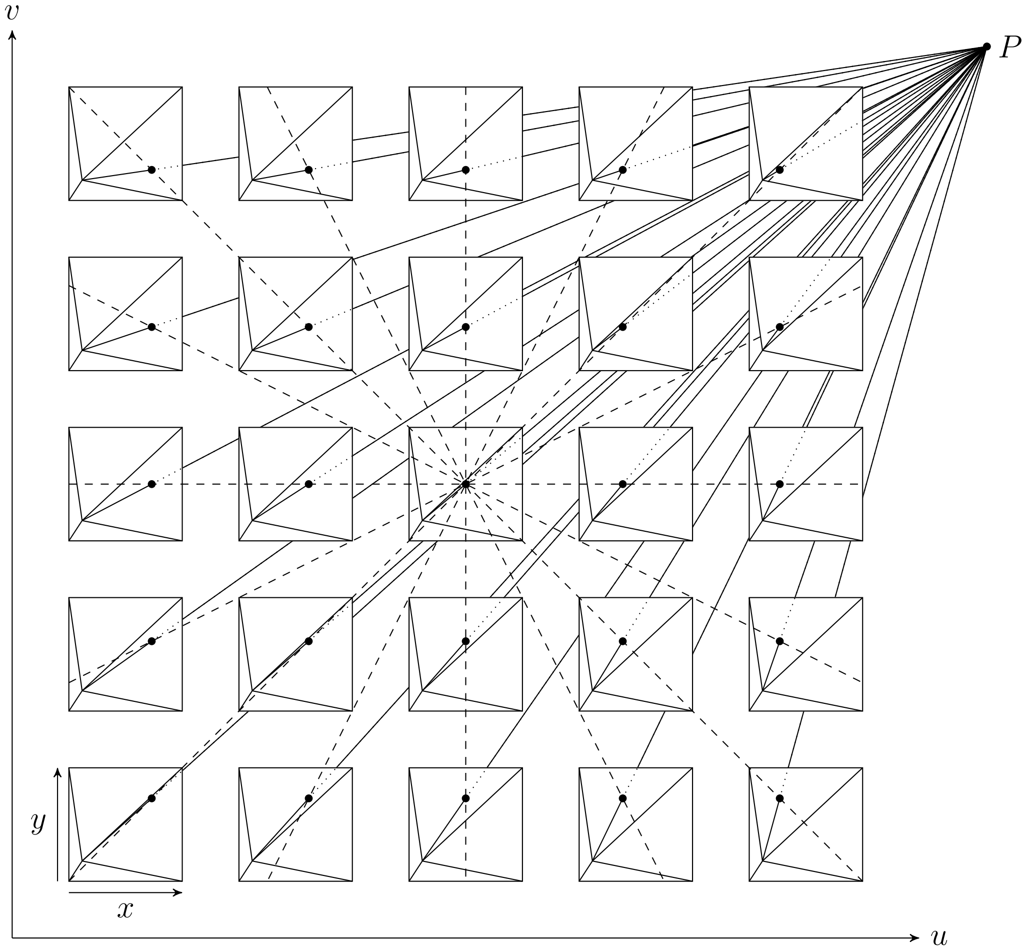


Figure 2.13: Rectified light field camera array consisting of identical pinhole cameras aligned on a regular two-dimensional grid. The view coordinates $(u, v)^\top$ are the grid positions of the cameras. The image coordinates $(x, y)^\top$ are the pixel grid positions on each camera sensor. Note that the projections of a three-dimensional point P, depicted as dots, lie on common epipolar lines, visualized as dashed lines.

to disparity d is expressed as

$$\frac{b}{Z} = \frac{b-d}{Z-f} \tag{2.5}$$

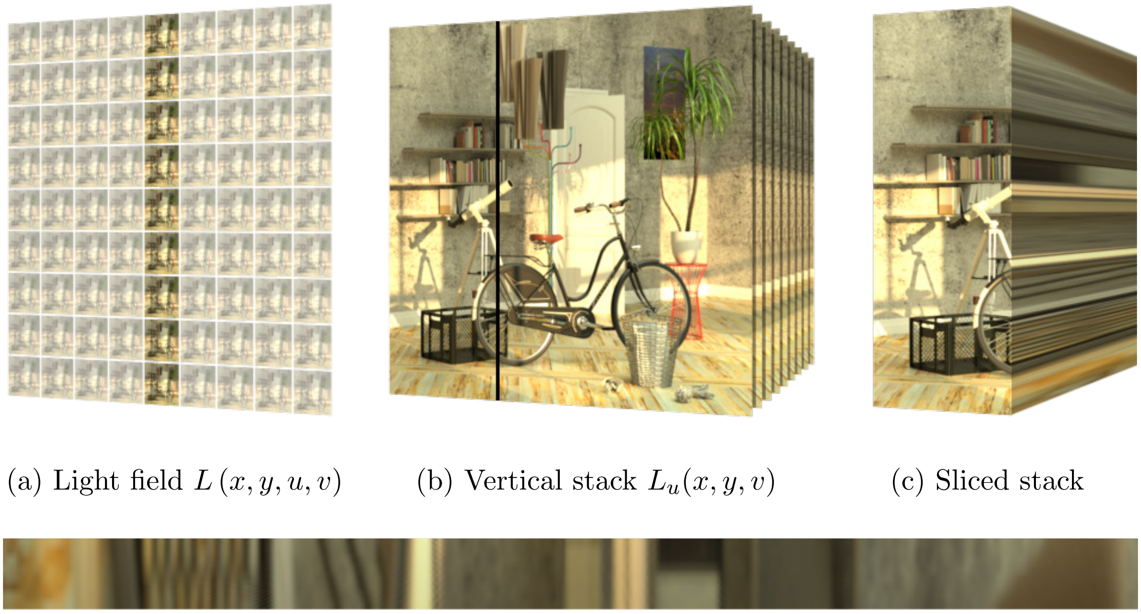
which simplifies to

$$Z = \frac{fb}{d}. \tag{2.6}$$

Finally, the world coordinates $(X, Y, Z)^\top$ of P can be calculated from p_0 and Z, utilizing the camera parameters.

Epipolar-Plane Images (EPIs)

For light field depth estimation, the introduced definition of epipolar geometry can be extended from a stereo camera pair to planar light field camera arrays. Figure 2.13 illustrates a rectified 5 × 5 light field camera array configuration. All cameras share identical intrinsics, are perfectly parallel, and are aligned on a regular two-dimensional grid. Similar to stereo rectification, recordings from imperfect



(a) Light field $L\left(x, y, u, v\right)$

(b) Vertical stack $L_u(x, y, v)$

(c) Sliced stack

(d) Vertical EPI $L_{x,u}\left(y, v\right)$

Figure 2.14: Extraction of an EPI. One exemplary column from all view points (a) is selected and stacked (b). An EPI is extracted, by slicing along one view dimension (d). Scene from Honauer *et al.* [10].

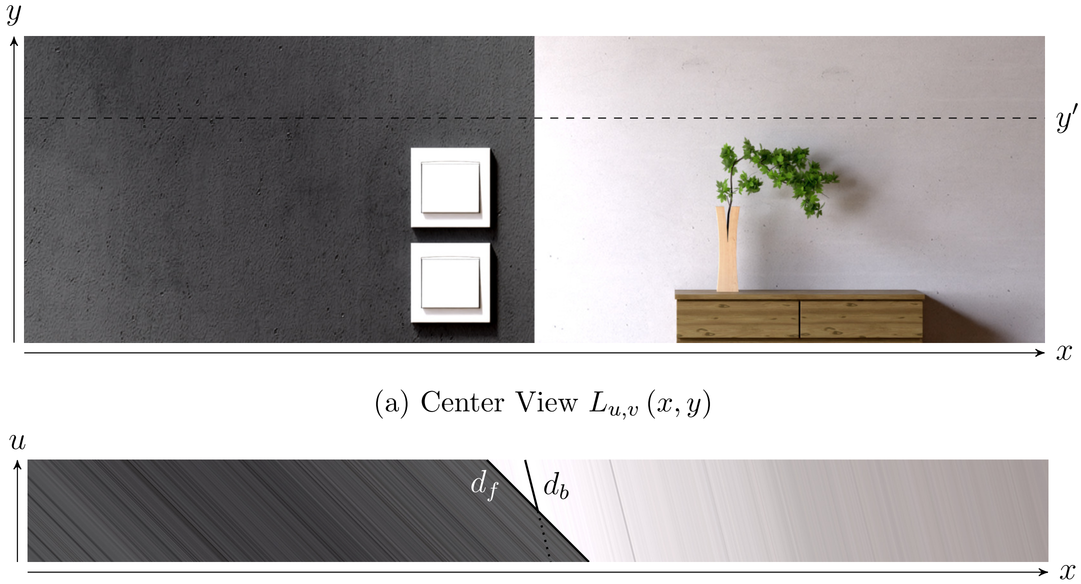
real-world camera arrays can be adjusted to meet these requirements. This can be achieved, for example, through the calibration and rectification method for light field cameras proposed by Schilling [69].

Analogous to stereo camera pairs, epipolar constraints exist between each pair of cameras within the rectified light field array. As depicted in fig. 2.13, the projection of a real-world point P remains constrained to epipolar lines. Similar to the standard rectified stereo geometry, trivial epipolar lines exist, aligning with the image rows y for a camera row v and with the image columns x for a camera column u. Additionally, non-trivial, diagonal epipolar lines can be established between any symmetric set of cameras, as demonstrated in fig. 2.13. The disparity can be generalized to the Euclidean distance between the projections p_0 and p_1 of a point P in two views, calculated as

$$d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}. \tag{2.7}$$

Given a known baseline b, which is the distance between the camera centers, the depth of a point P can be calculated similarly to the stereo camera pair using eq. (2.6). Although defined for camera arrays, this method is universally applicable to planar light fields captured by any method introduced in section 2.2.

For a given light field, one of the epipolar lines depicted in fig. 2.13 can be illustrated as an Epipolar-Plane Image (EPI). An EPI constitutes a light field captured along a specific line in image coordinates and the corresponding line in view coordinates. Two trivial examples are horizontal EPIs $L_{y,v}\left(x, u\right)$ and vertical EPIs $L_{x,u}\left(y, v\right)$, although EPIs can be derived across all angles within the light field. Figure 2.14 illustrates a vertical EPI, extracted from a 9×9 light field. This particular



(a) Center View $L_{u,v}(x, y)$

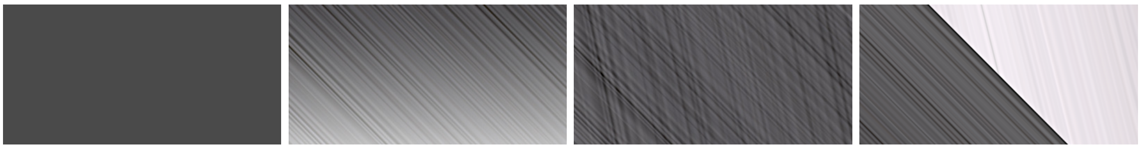
(b) Horizontal EPI $L_{y,v}(x, u)$

Figure 2.15: EPI of an idealized scene featuring two fronto-parallel walls. The darker wall, being closer to the camera, results in line structures with steeper slopes. Conversely, the white wall, positioned further from the camera, leads to less steep lines, which become occluded behind the closer wall.

EPI encompasses all views from the central column, as indicated in fig. 2.14a. To visualize the EPI extraction from these views, they are sequentially stacked as shown in fig. 2.14b and subsequently sliced along a designated image coordinate column x, as demonstrated in fig. 2.14c. Given that the stack comprises only nine views, it has been scaled and interpolated for better visualization. A portion of the resultant EPI is presented in fig. 2.14d. Note the presence of line structures with varying slopes. These lines are formed by corresponding points across all views. The disparity between correspondences in image coordinates determines the steepness of the slopes: larger disparities yield steeper slopes. More precisely, the slope of a line is equal to the negative inverse disparity of the points along that line.

Figure 2.15 visualizes this concept through an idealized real-world scene featuring two fronto-parallel walls, with the dark wall closer to the camera and the white wall further away. Figure 2.15b shows the line structures in the horizontal EPI at y'. Notice the steep slopes of the foreground dark line structures, which have a disparity d_f, in comparison to the brighter structures with a disparity d_b. Also note how the background structures are occluded by the foreground ones. The two lines in fig. 2.15b illustrate the two disparities visible at the occlusion boundary in the center view. Foreground points, like the edge of the front wall remain visible across all views of the light field. In contrast, the point on the back wall, which lies just behind this edge from the center view's perspective, is only visible in views to the right of the center. In the views left of the center view, the respective background line structure is obscured, illustrated by a dotted line. This results in the occlusion of parts of the background line structures.

Leveraging these properties, EPI-based light field depth estimation methods aim to accurately recover the slopes of line structures for each pixel. A straightfor-



(a) Non-textured (b) Specular (c) Semi-transparent (d) Occluded

Figure 2.16: EPIs, illustrating challenges for EPI-based light field depth estimation. Non-textured areas (a) introduce ambiguities. Specular reflections (b) disrupt the photoconsistency along line structures. Semi-transparent surfaces (c) blend with the background. Occluders (d) obscure parts of the background line structures.

ward and naïve method would compute the photometric consistency along the line structure for all possible slopes and selecting the most consistent one. Although this method could be effective in ideal scenes, real-world recordings present several challenges that complicate the task, as shown in fig. 2.16.

The most common challenge involves non-textured areas, depicted in fig. 2.16a. In such areas, photometric consistency remains the same across all possible disparities, leading to ambiguity. A common strategy to address this issue is to propagate disparities from adjacent textured areas, assuming local smoothness.

Specular reflections, shown in fig. 2.16b, pose another significant challenge. Most EPI-based depth estimation methods assume that scene objects are Lambertian, reflecting light uniformly in all directions. This assumption implies that each point in the scene distributes the same amount of light to every view in the light field. However, objects with strong specular reflections deviate from this model, reflecting light more intensely in certain directions. This leads to gradients in the EPIs, reducing photoconsistency despite correct disparity. The phenomenon of specular reflections in light fields has been extensively studied by Criminisi *et al.* [70], yet accurately predicting the disparity of specular surfaces remains a challenge, even for current state-of-the-art methods.

Semi-transparent surfaces are a challenging problem as well, as illustrated in fig. 2.16c. In EPIs, semi-transparent surfaces blend visually with those behind them, leading to ambiguities due to multiple visible line structures for a single point. Modeling semi-transparent surfaces and other cases of multiple valid depth modes is one of the subjects addressed in this thesis and will be discussed in detail in chapter 5.

Occlusions, highlighted in fig. 2.16d, present another common issue. Near the edges of foreground objects, the EPI contains line structures from the foreground visible across all views, as well as partial line structures from the background, visible only in certain views. This results in ambiguities and reduces photoconsistency for lines associated with the background object. Various approaches have been proposed to address occlusions, with some, like Schilling *et al.* [71], explicitly modeling them, and others, like Zhang *et al.* [72], aiming to avoid occlusions by selecting EPIs parallel to the occlusion boundary.

Addressing these challenges has been a subject of research for nearly four decades. Section 3.1.2 reviews some of the most significant contributions to the field.

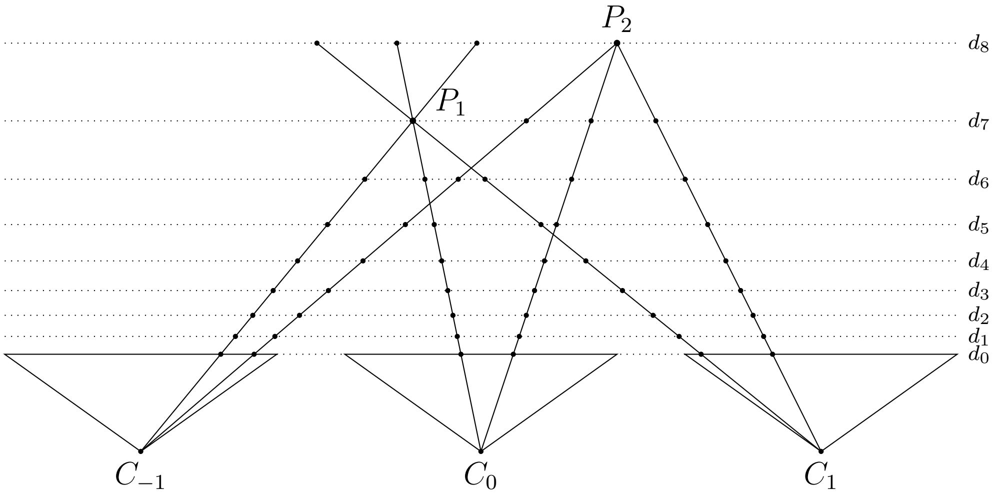


Figure 2.17: Plane sweep, demonstrated with three cameras. The projection of the three views onto a set of planes at discrete disparities $\{d_0, \ldots, d_8\}$ is illustrated for two points P_1 and P_2. Note that the rays passing through all images of a three-dimensional point converge at a specific plane, provided the point lies on that plane.

2.4.2 Plane Sweep

An alternative to using the epipolar properties of a rectified light field is the plane sweep concept, initially introduced by Collins [73]. The core principle involves sweeping a series of planes through the scene, with each view being projected onto these planes and assessing the photoconsistency among the projections.

Figure 2.17 illustrates three cameras C_{-1}, C_0, and C_1, and their projections onto a set of planes at the disparities $\{d_0, \ldots, d_8\}$. The projection is exemplified by two points P_1 and P_2. Note the images of P_1 in each of the views at d_0. By casting rays through these images, they are reprojected onto each plane. While the projections scatter across various points on all but one plane, the rays intersect at the correct plane, thereby reconverging all three images of P_1 to a singular point on the plane at d_7, aligning with P_1 itself. As a result, the color of all three projections at this point will be similar, yielding strong photoconsistency and serving as a depth cue. Conversely, the images of P_2 project to three distinct points on the plane at d_7, yielding poorer photoconsistency than the projections at the correct disparity d_8.

The primary advantage of this method lies in its flexibility, as it does not require rectified views. Instead, views can be projected onto the planes through a set of homographies, rendering plane sweep particularly suitable for non-planar light field camera configurations. Moreover, the view for which the disparities are estimated does not need be one of the actual views. Instead, the planes can be defined from the perspective of any virtual camera positioned within the scene, which makes it also suitable for image-based rendering, as showcased by Mildenhall *et al.* [74].

Should the views be rectified, the projection onto a plane can be accomplished simply by horizontally and/or vertically shifting each view, with the direction and displacement depending on the view coordinates and disparity of the target plane. This concept will be applied in chapter 4 to overcome the receptive field limitations of CNNs in predicting disparities from large-baseline light fields.

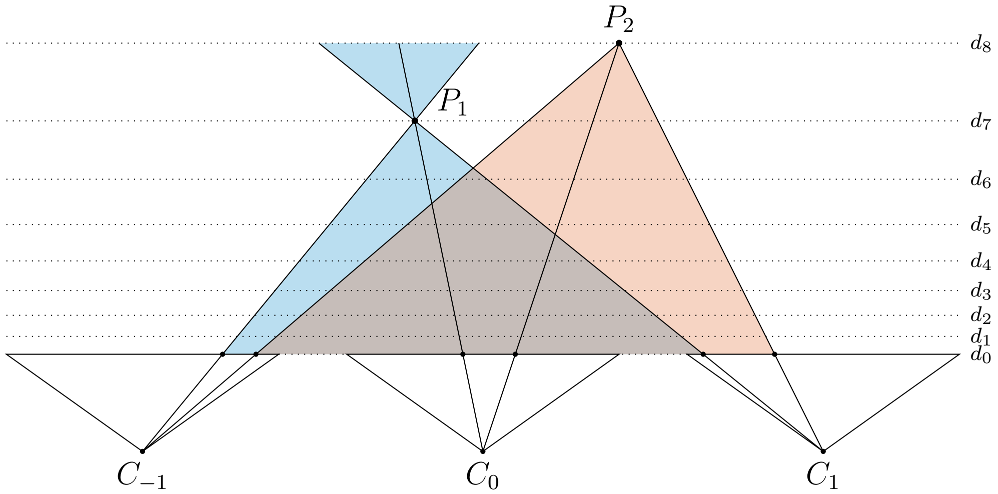


Figure 2.18: Depth from Defocus (DfD). The points P_1 and P_2 are in focus when the light field is refocused to their respective disparities d_7 and d_8. When refocusing to any other disparity, the points spread larger areas, visualized in blue and orange.

2.4.3 Defocus

Another depth estimation method is Depth from Defocus (DfD), which was briefly introduced in section 2.3.2. DfD estimates the amount of blur at each pixel. It is based on the idea that the circle of confusion increases with the distance of an object from the focus plane of the camera. However, one issue is that the amount of blur increases in both directions from the focal plane, leading to ambiguity. Traditionally, to overcome this, multiple images, known as a focal stack, are recorded by focusing at various distances using a conventional camera. Light fields allow for synthetic refocusing without the need to capture a focal stack. Figure 2.18 illustrates three cameras that are refocused to a set of planes at different disparities. A naïve method for light field refocusing projects all views onto a set of planes at $\{d_0, \dots, d_8\}$, akin to plane sweep. For each plane, the colors from all projected views are averaged to create one image in the focal stack. The blue and orange areas illustrate how the projections of two points P_1 and P_2 spread across the images in the resulting focal stack. Each of the points is only in focus on the plane that matches it's disparity.

Numerous methods have been proposed, some based solely on refocused light fields and others combining EPI-based methods with DfD. Most suggest that EPI-based methods are more accurate, while defocus-based methods offer more robustness, for example, at occlusions. By integrating both methods, it is believed that their strengths can be combined. However, the effectiveness of such combined approaches remains a subject of debate. Schechner and Kiryati [75] analyzed these claims, identifying a duality between the methods. This duality is evident when comparing fig. 2.17 with fig. 2.18. They noted that the assumptions largely stem from scale differences in camera setups, which are absent in light field refocusing. Therefore, the differences should theoretically diminish for refocused light fields, as both the focal stack and EPIs are derived from the same camera setup.

Section 3.1.3 discusses DfD methods specifically designed for light fields, and section 3.1.4 provides a summary of combined approaches.

2.5 Depth Estimation Methods

This section delves into the transformation of depth cues, extracted from light fields via previously detailed methods, into actual depth maps. It starts with an analysis of classical optimization-based techniques, followed by an overview over deep learning-based approaches.

2.5.1 Classical Methods

Classical methods for light field depth estimation have been in application since their inception by Bolles *et al.* [44] in 1987. These approaches have evolved alongside stereo depth estimation techniques, with both domains benefiting mutually. Contrastingly to stereo depth estimation, which relies on correspondence search between merely two views, light field depth estimation gains from considerably higher redundancy due to an increased number of views. This advantage is crucial for mitigating noise and proves particularly advantageous in the presence of occlusions.

A naiv̈e approach to light field depth estimation might solely depend on one or several depth cues, as introduced in section 2.4, opting for the most plausible disparity for each pixel. Nonetheless, this strategy can result in inaccurate estimates in scenarios where depth cues are unreliable, such as those depicted in fig. 2.16 for EPI-based estimation. A prevalent method to refine estimates involves the assumption of piece-wise smoothness within the scene, favoring relatively constant disparities in ambiguous regions over significant disparity jumps. Under this assumption, disparities can be propagated from neighboring regions with more certain predictions. An energy-minimization problem is commonly formulated based on these principles as

$$E(d) = E_d(d) + \lambda E_s(d), \tag{2.8}$$

comprising a data term E_d and a weighted smoothness term λE_s. The data term E_d evaluates the cost for each potential disparity based on one or multiple depth cues. The smoothness term E_s, varying significantly across methods, models the piece-wise smoothness assumption. While some methods model disparity as nearly constant within a local neighborhood, others, such as Baker *et al.* [76], depict it as a locally slanted plane. A recurring challenge is the suboptimal presumption of smoothness terms in proximity to occlusions. To address this, techniques such as the method by Schilling *et al.* [71] explicitly incorporate occlusions within the smoothness term, for instance, by employing a bilateral filter.

Energy minimization can be realized through various strategies. One category adopts a pixel-wise iterative method, like the algorithm introduced by Schilling *et al.* [71], which draws inspiration from PatchMatch Stereo [77] to iteratively extend certain disparity estimates to adjacent pixels. Conversely, alternative methods employ global optimization of the disparity map. For instance, Matoušek *et al.* [78] apply Dynamic Programming [79], Jeon *et al.* [80] and Lin *et al.* [9] employ GraphCut [81], and Bishop and Favaro [82] utilize gradient descent-based techniques. To improve the runtime, other methods like those by Heber and Pock [83] and Wanner and Goldluecke [84] leverage approximative approaches, and Neri *et al.* [85] implement a coarse-to-fine resolution pyramid.

2.5.2 Deep Learning-Based Methods

Since the rise of deep learning models in most computer vision domains over the past decade, these methods have been increasingly applied to depth estimation tasks. Remarkably, although deep learning approaches have dominated areas like image classification since the introduction of AlexNet [86] in 2012, it took until relatively recently for neural networks to surpass classical methods in light field and stereo depth estimation. EPINET [20], introduced in 2018, was the first method to match classical methods in performance on the HCI 4D Light Field Dataset [10]. In stereo depth estimation, deep learning methods have excelled in benchmarks such as KITTI [87], yet classical methods continue to lead in the Middlebury stereo benchmark [88].

Several explanations for this phenomenon exist. One perspective is that tasks like stereo correspondence search and EPI-based light field depth estimation might be inherently more suited to classical methods than complex tasks like image classification. They can be executed within a relatively small pixel window without the necessity for complex global pattern recognition, thus not benefiting significantly from learned prior knowledge. Another factor could be the lack of large, domain-specific training datasets. This is evidenced by the performance gap of CNNs between the KITTI [87] and Middlebury [88] stereo datasets, where the availability of extensive domain-specific training data for KITTI led to superior neural network performance. In contrast, the Middlebury benchmark, containing more diverse scenes, still suffers from a lack of depth training data. This pattern also applies to the HCI 4D Light Field Benchmark [10], which was not originally designed for deep learning and includes only 16 additional scenes that can be used for training. A third consideration is the fact that large, high-resolution images are still challenging for CNNs, mainly due to computational and memory constraints. This could also be a viable explanation highlighted by the performance disparity between the lower-resolution KITTI and the high-resolution Middlebury benchmarks. Moreover, high-resolution wide-baseline light fields present challenges due to the limited receptive field of CNNs, a topic addressed in chapter 4. Despite these challenges, deep learning has still become the leading approach in light field depth estimation, notable not only for its performance but also for the significant reduction in runtime, broadening the scope of potential applications.

Deep learning is applied in various ways for light field depth estimation. Initial methods, such as those proposed by Heber and Pock [89], [90], were still based on the classical energy minimization framework, employing a CNN solely to replace the data term E_d, with the depth map subsequently derived via an optimization process. Later approaches, like Shin *et al.* [20], directly predict the disparity map end-to-end, taking only a subset of light field views as inputs due to graphics memory limitations. More recent methods, such as those by Tsai *et al.* [21] and Chen *et al.* [91], use the entire set of light field views, recognizing that some views provide more valuable disparity cues than others. They therefore introduced an attention mechanism to prioritize information from different views accordingly. Some methods, paralleling classical approaches, use defocus cues instead of EPI cues. Zhou *et al.* [92] first proposed a deep learning method utilizing a focal stack synthesized from a light field and later developed a hybrid approach [93] that integrates cues from both the focal stack and EPIs.

2.6 Summary

Before reviewing related work, this section gives a brief summary over the chapter, highlighting key insights that will be essential in the following chapters.

Section 2.1 — The Plenoptic Function: Adelson and Bergen [23] introduced the plenoptic function, a mathematical model for the intensity of light in space. This five-dimensional model, capturing light from all directions at any point in space, is effective for vision tasks, especially light field photography. Yet, this model is not feasible for many applications, due to the redundancy of modeling empty space and the practical impossibility of capturing the entire function due to the theoretical camera's interference. A more feasible approach measures the plenoptic function within a limited volume of empty space, reducing its domain to four dimensions through the two-plane model, effectively used in light field cameras. This four-dimensional model is used throughout this thesis.

Section 2.2 — Light Field Photography: A planar light field camera is conceptualized as an assembly of identical pinhole cameras arranged in a regular two-dimensional grid. Camera arrays are a direct implementation of the light field camera model, capturing three- or four-dimensional light fields. Their advantage lies in capturing high-resolution videos of dynamic scenes through simultaneous camera triggering, although this necessitates high-bandwidth video processing and storage systems. Light field gantries provide a cost-effective alternative, using a single camera on a mechanized rig to methodically capture light fields of static scenes. While affordable and capable of producing high-resolution output, gantries are limited to static scenes and require elaborate mechanical controls and calibration. Plenoptic cameras offer a compact and accessible solution, utilizing a microlens array to capture multiple rays from each point in the scene. This technology captures four-dimensional light field information but involves complex analysis and calibration to process microlens array images into views. Developments in plenoptic camera technology highlight the balance between compactness, cost-effectiveness, and the inherent trade-offs in image resolution and quality. Coded aperture cameras address the resolution trade-off by using varying aperture patterns to control the light entering the camera, enabling the recording of detailed light fields without sacrificing spatial resolution. Despite their high-resolution capabilities, coded aperture cameras face challenges in light field quality dependent on the number of captured images and limitations of current hardware configurations. Synthetic light fields, generated from computer-rendered scenes, offer an alternative approach for creating light field data, useful for real-time rendering and training deep learning models for light field applications. Synthetic data is crucial for training algorithms in depth estimation from light fields, overcoming the lack of real-world depth capture technologies.

Section 2.3 — Overview of Depth Estimation Methods: Depth estimation methods in Computer Vision can be categorized into active and passive approaches. Active techniques use artificial light for scene illumination, while passive methods rely on the existing light in the scene. Active methods include Structured Light, projecting a pattern onto the scene and calculating depth from pattern deformations via triangulation, ToF sensors, measuring the time light takes to return after hitting objects to generate a depth map and LIDAR, using laser beams to create a three-dimensional map based on the time for beams to bounce back. Passive methods comprise Monocular Depth Estimation, inferring depth from a single image using visual cues, SfM analyzing camera movement or multiple images to deduce scene structure. DfD, estimating depth by analyzing image blur levels, stereo depth estimation, estimating depth from disparities between two images and MVS using multiple cameras to triangulate scene points.

Section 2.4 — Sources of Depth Cues: Sources of depth cues in light fields include epipolar geometry, plane sweep, and defocus cues. Epipolar geometry is essential for depth estimation, utilizing triangulation to infer depth from disparity. It simplifies finding corresponding points between images captured from different viewpoints. Rectification aligns epipolar lines with image rows or columns, easing the correspondence search. Epipolar geometry extends to light field camera arrays, simplifying correspondence search with epipolar constraints. EPIs visually represent these constraints, showing disparity as the slope of line structures. However, challenges like non-textured areas, specular reflections, semi-transparent surfaces, and occlusions complicate EPI-based depth estimation. Plane sweep offers an alternative method, projecting views onto planes at different disparities and assessing photoconsistency. This technique adapts to non-planar light field configurations and doesn't require rectified views, offering depth estimation flexibility from various perspectives. DfD measures blur at each pixel to infer depth, utilizing the circle of confusion's variation with distance from the focus plane. Light fields allow to generate a focal stack by synthetic refocusing. This method is considered more robust in scenarios like occlusions, though integrating it with EPI-based methods for improved accuracy remains debated.

Section 2.5 — Depth Estimation Methods: Classical depth estimation methods predate deep learning and are often based on some form of optimization. These approaches often assume piece-wise smoothness in scenes to refine depth estimates, using energy minimization that combines data and smoothness terms. Challenges include handling occlusions and the variability in smoothness assumptions. Energy minimization techniques range from pixel-wise iterative to global optimization, employing methods like Dynamic Programming, GraphCut, or gradient descent. Some strategies reduce computation time through approximations or multi-resolution approaches. With deep learning's rise, it has increasingly been applied to depth estimation, eventually surpassing classical methods in certain benchmarks. The delay in deep learning's dominance is attributed to the specific suitability of classical methods for tasks like correspondence search, the lack of large, specific datasets, and the challenges of processing high-resolution images. Still, deep learning has led to significant performance and runtime improvements in light field depth estimation. Initial deep learning applications focused on replacing components of classical frameworks, progressing to end-to-end disparity map prediction. Recent methods utilize the full set of light field views, with attention mechanisms to prioritize valuable disparity cues. Some methods also integrate defocus cues from synthesized focal stacks with EPIs for comprehensive depth estimation.

3. Related Work

This chapter provides an overview over previous contributions that are essential to this thesis. It starts with an exploration of light field depth estimation, detailing the progression of techniques from classical methods to advanced deep learning approaches. The discussion then transitions to light field rendering, with a focus on depth inputs, a crucial ingredient for good rendering results. Lastly, the chapter addresses posterior regression with neural networks, which is relevant for multimodal depth estimation, introduced in chapter 5.

3.1 Light Field Depth Estimation

This section offers an overview of the development of light field depth estimation methods. Figure 3.1 shows the number of publications related to light field depth estimation per year since 1987. The considered publications are: [2], [6]–[11], [20], [21], [33], [44], [56], [66], [70]–[72], [78], [80], [82]–[85], [89]–[289]. Interest in the area has noticeably grown in recent years, particularly over the last decade. This increase likely stems from the availability of plenoptic cameras to consumers. In the past four years, the publication count has slightly decreased and plateaued. This decrease might be due to the commercial failure of plenoptic cameras for end users or possibly a short-term drop in publications caused by the COVID-19 pandemic. Despite these factors, interest remains high, reflecting the growing importance of light field cameras in industrial uses.

This section initially highlights surveys that have been conducted on this topic. The methods in light field depth estimation are then categorized into four primary groups: First, EPI-based methods, which focus on analyzing the unique geometric properties of light fields. Second, defocus-based methods, which rely on synthetically refocused or defocused renders of the light field to estimate depth. Third, methods that combine both EPI-based and defocus-based approaches, aiming to combine the strengths of both methods. Fourth, approaches that rely on deep-learning, which are the most recent contributions to the field. Finally, the section explores the most widely used datasets and benchmarks in the scientific community.

3.1.1 Surveys

Two surveys have been conducted on the topic in 2017 and recently in 2023. Both compared the state-of-the-art algorithms at the time and analyzed remaining challenges.

Johannsen *et al.* [137] provide an overview of participants in the second workshop on Light Fields for Computer Vision. They categorize all 14 methods presented ac-

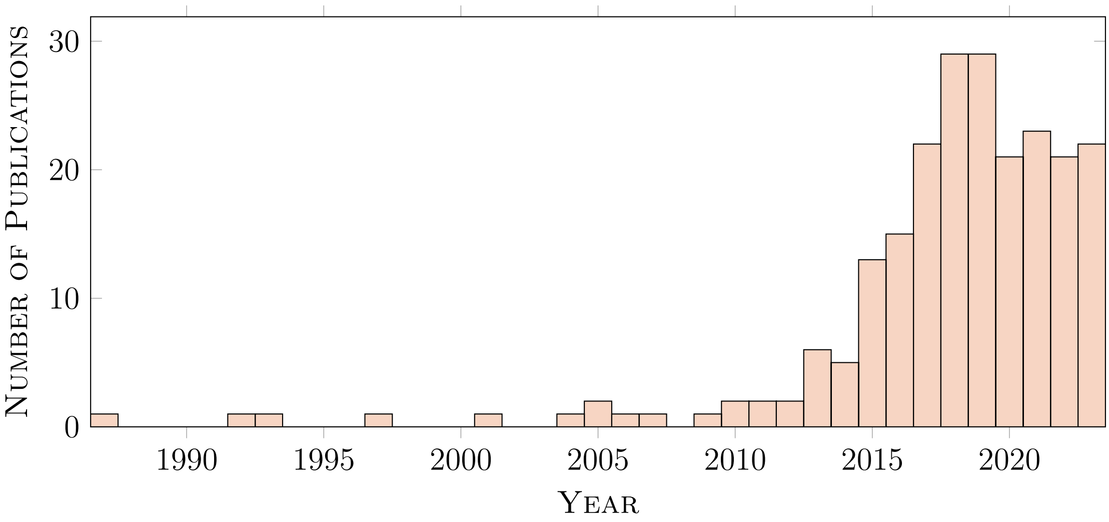


Figure 3.1: Number of publications on light field depth estimation by year. The interest steadily increased in recent years and still remains high.

cording to their light field representations. These categories include methods based on EPIs, view patches, the focal stack, and multi-view stereo. Their analysis encompasses initial data terms, optimization algorithms, and refinement post-processing used in each method. All algorithms undergo evaluation against a broad array of metrics. These metrics assess occlusion handling performance, robustness to errors in input views, and more. A key finding from their study is that no single algorithm excels in every evaluated category. Their research indicates that, as of 2017, while the overall performance of algorithms is strong, challenges persist in occlusion modeling and runtime efficiency.

Wang *et al.* [2] conducted a comprehensive survey of 38 methods for light field depth estimation. Methods are categorized into traditional and deep learning-based, with further subcategories. Traditional methods include those based on EPIs, stereo matching, and the focal stack. Deep learning-based methods are categorized into those using EPIs, stereo matching, and other structures. The survey also examines novel approaches like unsupervised methods, Non-Lambertian methods, and depth estimation for all views. All methods were evaluated against metrics such as robustness to input noise and occlusion. The article identifies future research directions, including improving training data, optimization algorithm runtime, interpretability, practical application integration, and generalization to unknown data.

3.1.2 Based on Epipolar Plane Images

The analysis of EPIs is probably the most inherent method for depth estimation from light fields. In this section, classical methods that use EPI properties for depth estimation in light field images are explored. Table 3.1 classifies these methods by input, output and design goals.

In 1987, Bolles *et al.* [44] pioneered the topic of depth estimation from light fields. The method they proposed involves using a moving camera mounted on a robot, similar to a translation stage, to render a three-dimensional image of a static scene. This groundbreaking work introduced and analyzed EPIs for the first time. The authors discovered, described, and demonstrated that linear local features in

Year	Method	Input			Output			Design Goals		
		3D	4D	Plenoptic	Dense	Sparse	Segmentation	Occlusion	Specular	Runtime
1987	Bolles *et al.* [44]	x				x				
1997	Werner *et al.* [94]	x				x				
1998	Baker *et al.* [76]		x		x		x			
2001	Matoušek *et al.* [78]	x			x					
2005	Criminisi *et al.* [70]		x		x		x	x	x	
2006	Berent and Dragotti [290]	x					x	x		
2010	Bishop and Favaro [98]			x	x					
2011	Wanner *et al.* [102]			x	x				x	
2013	Wanner and Goldluecke [84]		x		x					x
2013	Kim *et al.* [291]	x			x					x
2014	Heber and Pock [83]			x	x					x
2014	Chen *et al.* [108]		x		x			x	x	
2015	Jeon *et al.* [80]			x	x			x		
2015	Neri *et al.* [85]			x	x			x		x
2016	Zhang *et al.* [72]		x		x			x		
2016	Johannsen *et al.* [123]			x	x				x	
2017	Sheng *et al.* [147]		x		x			x		
2018	Sheng *et al.* [177]		x		x			x		
2018	Jeon *et al.* [165]			x	x			x		
2018	Schilling *et al.* [71]		x		x			x		

Table 3.1: Classification of EPI-based methods by input light field type, output depth type, and method focus. Inputs are either a three-dimensional light field with one view dimension or a four-dimensional light field with two view dimensions, with some methods tailored for light fields recorded with plenoptic cameras. Outputs range from dense depth maps and sparse depths to segmentation into larger areas with similar depth. Some methods focus on performance at occlusions or specular reflections, while others are optimized for faster runtime.

an EPI correspond to disparity values. To extract these linear features, their algorithm detects edges, peaks, and troughs in the epipolar geometry. This process involves taking the second derivative along the image dimension and identifying its zero-crossings. Lines are then fitted along these features and merged with collinear detections to remove outliers. While computationally expensive, this intuitive algorithm can detect local disparity values for objects. A *free-space-map* is formed for each EPI-based on all sparse detections. In the final step, the information is linked across all EPIs. The results obtained were sparse and limited to highly structured areas due to inaccurate calibration and the low frame-rate of the imaging sensor used at the time. Over the following decades, light field imaging hardware has undergone significant advancements, resulting in smaller and more accurate sensors and camera arrays. This progress has led to constantly improving depth accuracy, making it a more viable depth sensing solution.

Werner *et al.* [94] introduced a method to render novel views from three-dimensional light fields, which are recorded on rotation or translation stages. In their argument, they assert that stereo matching does not suffice to establish correspondences due to ambiguity, hence the necessity for more views. Edge features are detected in all views utilizing the one-dimensional Deriche detector. Subsequently correspondence matching between two neighboring views is performed: Starting from the first view, correspondences between the edge features of two adjacent views are matched. By iterating over all the views, a chain of correspondences is established.

As a result, correspondences between two distant views are preserved via this chain of pairwise correspondences. They demonstrated novel view rendering using their proposed method: To project all pixels to a novel view in-between recorded views, they utilize linear interpolation between two adjacent sparse edge features. The results indicate that their method is more effective when there is a significant distance between views. Nonetheless, it remains unclear if this only applies to rotation or translation as well. Furthermore, the correspondences are sparse and rely heavily on a large number of strong edges present in the light field, consequently, only simple scenes comprising a single object were demonstrated.

Baker *et al.* [76] focus on rendering novel views derived from multi-view inputs, but also contribute a method for depth estimation. The authors propose using a *layered scene representation*, today often also referred to as Multi-Plane Images (MPIs). Each layer in this structure is approximately planar, with the slant of each layer defined by an explicit plane equation. Additional per-pixel depth offsets improve details and per-pixel transparency enables realistic stacking of semi-transparent layers. To achieve this, a two-step approach is proposed: First, opacity is ignored, each pixel on a layer is deemed as either visible in an image or not, in order to establish an initial scene structure. Segmentation into layers needs to be manually initialized by the user. Using an optimizer, plane equations are estimated in such a way that the homographies between two images warp the pixels on the layer consistently across multiple views. Color values are assigned to each layer, similar to an image texture, by blending warped images from all views onto the layer. To estimate per-pixel depth offsets, a stereo matching algorithm is employed for each plane separately. Subsequently, more pixels adjacent to already assigned pixels are added from all views to each layer based on photoconsistency. The second step involves refining the layers and inferring opacities. A generative model is established, where all layers are warped to each image and alpha-blended from back to front using the *over* operator. Leveraging this generative model, gradient descent is utilized to refine pixel colors and opacities. Subsequently, the plane equations and depth offsets are iteratively recalculated as described above. The authors conducted experiments on two multi-frame datasets, which show good results for novel view rendering. However, the requirement of human input for initial layer segmentation presents a labor-intensive process. In addition, only Lambertian reflection is assumed and the depth offsets seem to be inaccurate for the results presented.

Matoušek *et al.* [78] contributed the first purely signal-based approach for depth estimation from light fields. The authors raise concerns about the limitations of correspondence-based approaches due to missing correspondences in certain views, while pointing out that more local dense methods suffer from low accuracy. Consequently, they introduce a dense optimization-based approach for three-dimensional light fields captured with a translation stage. They model the disparity along one line in the light field as a continuous mapping from the position along the line to the respective disparity. As a cost term, the color variance along each disparity slope is chosen. Furthermore, the total cost function is constructed in a way that makes the cost independent of the disparity for non-textured areas. This construction is based on the hypothesis that an optimizer will smooth out non-textured regions. To optimize the total cost for each EPI, dynamic programming is used. The search domain is thus discretized on a regular mesh, and the disparity-curve along the EPI is modelled as a piecewise-linear function. Their methodology was tested successfully on a simple scene composed of two flat surfaces forming a common edge, similar to one edge of a cube, covered with a strong granite-like texture. Nevertheless, their approach is constrained by two assumptions: the data term only models Lambertian reflection and the disparity, modeled as a continuous function, only works for convex objects and thus, does not account for occlusions. Therefore, they conclude that their method is best suited as a component of a larger, more comprehensive solution.

Criminisi *et al.* [70] deal with depth estimation as well as specular removal from three-dimensional light fields captured by a linear translation stage. Their methodology involves dividing the scene into interconnected smooth regions, they term *EPI tubes*. Each two-dimensional EPI tube consists of a number of *EPI strips*, which are smooth one-dimensional lines or columns in the light field. To extract EPI strips, the authors proposed two strategies: The first strategy favors those parts of the scene, exhibiting minimal variation along a disparity line in the EPI strip. This is based on the idea that color variance is distorted if one scene element occludes another. Hence, the occluding scene element is chosen over the occluded one. Subsequently, the method fills gaps between extracted EPI strips. Once these strips are removed from the light field, the entire procedure is repeated until the whole light field is covered. Their second strategy, a refined version, begins by making educated predictions about the EPI strips in each row of the light field. It employs the Canny edge detector [292], based on the assumption that object borders lead to a visible color change along a line in the EPI. Each of the detected lines becomes a candidate for an EPI strip border. The cost of all potential strips is calculated, adjacent strips are merged, and the configuration with the lowest overall cost is chosen. This process is repeated, and any remaining gaps are filled until the entire EPI has been segmented into EPI strips. Although the article doesn't discuss how to merge EPI strips into EPI tubes, visible streaking artifacts in the results imply this might be a challenging task. In addition to depth estimation, the article also discusses the behavior of specular, or mirror-like, reflections in light fields. They discovered that the apparent location of a reflected object is located on a caustic surface which is influenced by the shape, orientation, and depth of the reflector. This leads to a phenomenon, they term *epipolar deviation*, meaning the specular reflection of a point deviates from the line in the epipolar geometry where Lambertian reflection is expected. However, they argue that in most scenes, this deviation is relatively small and conclude that it's therefore still feasible to differentiate between different types of reflections with EPI analysis. Further exploration led them to identify six types of specular reflections, contingent upon the texture of the reflecting and reflected objects. To summarize: textureless objects can create ambiguity unless the reflected object is a single point. If a specular textureless object is part of a correctly extracted EPI strip, its Lambertian color can be determined. However, a consistent ambiguity exists: the reflecting object might be a single entity, or possibly two entities, with the potential reflected object also being Lambertian and laying between the two entities. For specular removal, the authors suggest a method based on the observation that the color of a specular pixel changes along a disparity slope. They propose using the lowest intensity color value for each pixel as an upper limit for the Lambertian light component and the remaining difference as its specular component. This method was validated using a real scene and was successful in recovering some of the specular and diffuse components, although it introduced larger artifacts. To summarize: Criminisi *et al.* [70] laid substantial theoretical groundwork for estimating depth and material properties from light fields. While the presented algorithms yield to visible artifacts for depth estimation and specular removal, the article explores the characteristics of light reflection within the epipolar geometry with remarkable detail.

Berent and Dragotti [290] adopt the level set methodology to segment three-dimensional light fields into EPI tubes. Their objective is a relatively unsupervised segmentation technique anchored on disparity. In the proposed optimization scheme, occlusions and disocclusions of a point are modeled explicitly. Each EPI tube is represented by a contour, modeled as the zero level-set of a speed function. An iterative scheme based on least squares is employed to optimize occlusions and disocclusions: one tube is kept fixed while others are propagated and vice versa. The motivation behind this approach is to prevent errors in the segmentation of foreground tubes from propagating to occluded background tubes. When occlusions occur, the occluded tube is partitioned into two or more sub-tubes. Experimental results indicate the method's capability to accurately segment synthetic Lambertian fronto-parallel light fields. In addition, the authors demonstrate the segmentation of a real-world scene into discrete disparities with approximate correctness. Nevertheless, the method carries an inherent assumption that each EPI tube is roughly fronto-parallel and Lambertian. This assumption introduces limitations for most applications, as evident in the presented real-world results.

Bishop and Favaro [98] address the problem of depth estimation in plenoptic cameras, which differ significantly from camera arrays. Plenoptic cameras present several advantages including compactness and low complexity. However, a significant disadvantage is the low resolution caused by the shared image sensor, leading to images being undersampled for many depths and rendering texture matching ineffective. A previous solution to this problem involved filtering out aliased signal components and only matching the low-pass part which limits resolution. To solve this issue, the authors introduce depth-dependent filtering. This approach allows depth to be reconstructed at full sensor resolution. The authors also address the issue of missing data in the corners of each microlens image, a result of the circular main lens, using inpainting techniques for data completion. A self-consistency criterion is additionally employed to penalize discontinuities in the reconstructed full-resolution views. Depth estimation is achieved through energy minimization, using the conjugate gradient method, with a total variation term used for regularization on the depth map. Experiments are conducted on both synthetic and real data. While the results show promise, there are observed deficiencies including holes and blur artifacts.

Wanner *et al.* [102] introduced the pioneering method for extracting EPIs from light fields captured by focused plenoptic cameras. The primary challenge lies in the fact that these cameras do not directly capture views with a full depth of field, which is essential for forming sharp EPIs. The proposed approach comprises three steps: Firstly, full depth of field views are rendered. To accomplish this, the authors introduce a method that synthesizes views by integrating patches from microlens images, with a specific patch size. Given that the optimal patch size is locally dependent on the correct focal point, which is unknown if no depth information is available, the authors suggest an alternative method to estimate this patch size. Their estimation process involves minimizing the gradient magnitude at the borders of the image patches, ensuring smooth transitions and thereby reducing plenoptic artifacts. The estimated patch sizes can then be used to render full depth of field views. Secondly, views from different types of microlenses are merged. This step is required if the plenoptic camera incorporates various microlens types and can be omitted otherwise. Finally, EPIs are extracted from the resulting rendered views. Through experimental analysis using light fields recorded with a Raytrix camera, the authors demonstrate that their focal maps are superior to results from sparse stereo matching methods. They also showcase sharp full depth of field renderings and compelling EPIs, which can be leveraged for downstream depth estimation. Given that the local foci within the scene are dependent on the depth of each point, the introduced method may also be considered an implicit method for depth estimation. This initial rough depth map could then be refined by a downstream method, using the extracted EPI.

Wanner and Goldluecke [84] have put forth a method that computes the disparity directly via a structure tensor. This is grounded in the principle that disparity can be derived from the slope of lines in an EPI. An adapted structure tensor is applied to the four-dimensional light field to estimate the line slope. The tensor, encompassing all combinations of partial derivatives, is integrated over a given area utilizing a Gaussian window function. The strongest eigenvector of this tensor aligns with the EPI gradient. In textured regions, this eigenvector offers a precise disparity estimate. However, in non-textured areas and due to noise, the accuracy of the eigenvector diminishes. The method also tends to produce outliers near occlusions due to the presence of varying local features in these areas. Addressing these issues, the authors introduce a global optimization approach grounded in functional lifting and discrete depth labels. In addition to this, they propose a rapid denoising scheme utilizing $\mathcal{L}_1$-smoothing as a less costly alternative to global optimization. Experimental findings suggest that this approach results in inaccuracies in non-textured regions and outliers at occlusions. Yet, owing to its simplicity and robustness, the structure tensor still remains a prevalent tool for light field depth estimation.

Kim *et al.* [291] address the problem of *Gigaray* light fields, consisting of hundreds of views, where existing algorithms don't work efficiently due to the substantial amount of data and therefore memory consumption. Their contributions include computing accurate depths specifically around object boundaries rather than interior regions and then propagate smoothly into these regions. This operation is performed on individual light rays, bypassing the need for global optimization. In a first step to compress the light field, a more efficient representation is introduced, storing only a singular color value and the corresponding disparity for each column in the EPI. In a second step, pixels that are not in close proximity to an edge are rejected using an edge confidence metric. Disparity is computed for each edge pixel, using a modified Parzen window estimation with an Epanechnikov kernel. Next, the depth is propagated to non-textured and specular regions using a *fine-to-coarse* scheme: Higher resolution depth values are blurred and scaled down, which allows for filling adjacent pixels. Finally, the depth image is upscaled, filling every pixel with the disparity that was initially assigned to it. The authors present successful results on three-dimensional light fields with 100 views, captured with a Digital Single-Lens Reflex (DSLR) camera. However, the disparities obtained are still noisy and lack details.

Heber and Pock [83] introduce the concept of Robust Principal Component Analysis (RPCA) to light field depth estimation. It's based on the idea to warp all views to the center view based on per-pixel disparity. In their model, each warped view, with a warping factor depending on the unknown per-pixel disparity, corresponds to a column of a matrix. They utilize Principal Component Analysis (PCA) to minimize the rank of the matrix, splitting the light field into a low-rank component and noise, with the lowest rank implying true disparity across all pixels. An issue arises with traditional PCA, which is highly sensitive to sparse errors of high magnitude, such as occlusions and specular reflections in the context of light fields, or aliasing typical for plenoptic cameras. The proposed method allows for sparse errors of high magnitude, achieved by splitting the matrix of warped images into a low-rank component and a sparse component. This can be interpreted as a simultaneous optimization of matching and denoising. Given that this problem is Nondeterministic Polynomial Time (NP) hard, the authors have derived a convex approximation. Both synthetic and real-world experiments are presented for depth estimation and denoising. Despite the presented approach offering a simple and flexible solution to the problem, it remains vulnerable to occlusions and the optimization process is inefficient.

Chen *et al.* [108] developed a method specifically designed to handle heavy occlusions. Their approach is rooted in view statistics, specifically leveraging the Surface Camera (SCam) model [293]. The SCam image of a three-dimensional point consists of all pixels in the light field views that cast a ray through this three-dimensional point. In cases where the SCam lies on a Lambertian surface within the scene that is not occluded, all pixels of the SCam image are equally colored. For specular surfaces, the color varies slightly. However this variation remains negligible for low-baseline light field cameras, as demonstrated by Criminisi *et al.* [70]. When dealing with a partially occluded point, the SCam image is only partially constant, and when the occluded surface is non-textured, the SCam image remains the same across a larger depth interval. Addressing this issue, the authors introduce a consistency metric based on the bilateral filter. Their findings show that the ground truth depth of a pixel always represents a local minimum of their bilateral metric. To infer dense depth, they propose a parallel algorithm where all local minima are extracted and a confidence measure is calculated. If the confidence measure surpasses a predefined threshold, the global minimum is assigned as the pixel's depth. Otherwise, the depth of adjacent pixels is propagated. Experiments on synthetic and real data show, that the method proves particularly effective in situations with heavy occlusions, although its performance declines in large non-textured areas.

Jeon *et al.* [80] focus their research on plenoptic cameras. Unlike camera arrays, microlens recordings are known to experience additional distortion, noise, and blur effects. In response to this, the authors present an optimization-based distortion estimation and correction technique. The actual process of depth estimation is then executed through a pixel-wise cost volume and discrete feature matches. To accommodate the small sub-pixel shifts observed in narrow-baseline plenoptic camera recordings, the authors employ a phase shift grounded in discrete Fourier transform to generate this cost volume. The cost function is formed from two complementary functions: the Sum of Absolute Differences (SAD) and the Sum of Gradient Differences (GRAD). An edge preserving filter is then used to refine non-textured regions in the volume while maintaining sharp occlusion edges. In addition to the cost volume, discrete feature matches are estimated based on Scale-Invariant Feature Transform (SIFT). These two components, the dense cost volume and the SIFT features, are then used to globally optimize discrete depth labels, achieved through multi-label optimization with Graph Cut. The method concludes with a refinement step that iteratively transforms the discrete labels into a continuous depth. While the methodology exhibits promising performance on real images from plenoptic cameras, it does encounter issues with noise and artifacts at occlusion boundaries.

Neri *et al.* [85] proposed a method specifically designed for plenoptic cameras, which employs a multi-resolution optimization strategy. Their method uses discrete depth samples that are used to map a pixel coordinate from the center view to any arbitrary view within the light field. An energy measure for each pair of views is computed by averaging the difference of the $\mathcal{L}_2$ norms of the pixels neighborhoods. For non-textured areas, an extra smoothness term is added to the model. The model appropriately addresses occlusions by computing the energy twice for each side of a possible occlusion, with the minimal energy being selected. In a local optimization procedure, the depth that results in the minimum sum of energies across all pairs of views, is selected. To reduce runtime, a *coarse-to-fine* scheme on discrete depth labels is utilized. Upon calculating the energy for each discrete label, a subsequent refinement step selects continuous depth values. This step is conducted iteratively from low to high resolution, providing feasible candidates even for smooth regions in the process. Experimental results on synthetic data indicate that the method, while effective, still generates artifacts in large non-textured areas and at occlusion boundaries.

Zhang *et al.* [72] present a method based on the Spinning Parallelogram Operator (SPO), with a focus on its performance near occlusion areas. SPO divides a parallelogram-shaped region of the EPI into two distinct regions. The spinning parallelogram is parameterized via a weighting function, which adjusts the slope of its edges and the separation border according to the potential depths. The authors assume that the true disparity corresponds to a maximum χ^2 difference in the color distributions of the two regions. Consequently, at occlusions, the distance retains a local maximum for both potential line directions: the depth of the foreground surface and the depth of the occluded surface. Low-variance distances in ambiguous areas yield a low difference. Considering that different occlusions result in different estimates for horizontal and vertical SPOs, the authors apply a filtering method to integrate both estimates. To populate textureless areas, confident estimates are propagated to non-textured regions using a guided filter. Experiments conducted on both synthetic and real data provide promising outcomes. Nevertheless, the performance at occlusions remains dependent on the relative angle between EPI extractions and object boundaries.

Johannsen *et al.* [123] propose to extract multiple depth layers from a light field, resulting either from semi-transparent surfaces or specular reflections. They construct a dictionary of small EPI patches, or atoms, that facilitate a sparse local encoding of the light field. The underlying approach is founded on $\mathcal{L}_1$-sparse coding, also known as Lasso. Each atom corresponds to a distinct depth and is formulated by lifting parts of the light field. Consequently, encoding the light field using the dictionary also yields dense depth estimation. When multiple atoms contribute to a specific region, this suggests the presence of multiple valid depth layers within that area. To build the dictionary, three methods of varying precision and computational costs are introduced: The first method involves lifting a one-dimensional base patch, meaning a column or line from one input view, to a two-dimensional EPI patch. The second method consists of lifting a cross-hair patch using a horizontal and vertical slice from the light field, enforcing disparity consistency between those dimensions. The third method lifts a two-dimensional patch from the center view to a four-dimensional light field patch, offering the highest precision but at great computational cost. After solving the $\mathcal{L}_1$-sparse coding, a mask is created to divide regions with single valid depths from regions with two valid depths. Non-textured areas undergo inpainting using $\mathcal{L}_1$-inpainting with a weighted second-order Total Generalization Variation (TGV). Experiments conducted on both synthetic and real-world data, captured with a plenoptic camera, demonstrate promising results. Nonetheless, even with the capacity to estimate multiple modes, the presented method lags behind the overall performance of state-of-the-art deep learning-based frameworks.

Sheng *et al.* [147] also tackle the problem of occlusion estimation. Similar to the SCam model [108], their work is based on an Angular Sampling Image (ASI) that consist of all points in the light field views that are projected from a particular three-dimensional point within a scene. For a point on a non-occluded Lambertian surface, the ASI exhibits a singular, consistent color. However, if a point is occluded, certain portions of the ASI contain the color of the occluding surface. The authors establish that if the depth of the occluding surface is fronto-parallel, the ASI's occlusion boundary is similar to the occlusion boundary in the light field's center view. Leveraging this assumption, the authors formulate an integral guided filter to predict occlusion probabilities. Notably, this filter relies solely on the light field views, making it integrable into a variety of depth estimation frameworks. As a practical demonstration, the authors integrate the filter into a cost-volume based stereo matching framework. Experimental evaluation on both synthetic and real data illustrates that the integral guided filter enhances the definition of occlusion boundaries for a wide range of depth estimation methods.

Sheng *et al.* [177] proposed another method emphasizing performance near occlusion regions. Many previous techniques employ only a single or two orthogonal EPIs, an approach that has a distinct shortcoming: Both EPIs incorporate elements of the occluding surface at edges that are not perfectly horizontal or vertical. To counter this effect, the authors presented a strategy using an extended set of EPIs extracted at varying angles. Additionally, they put forth a novel depth estimation framework that starts by identifying potential occlusions. The algorithm exploits the variance in depth estimates across different EPI orientations for occlusion reasoning. In areas of occlusion, this variance increases due to the intermittent presence of the occluding surface. An initial depth map, extracted using the SPO method by Zhang *et al.* [72], serves as an additional indicator for occlusions. Using the derived direction of the occlusion boundary, the optimal non-occluded EPI is chosen to estimate the final depth value. In order to determine this orientation, the authors introduce the compass operator: This operator is akin to the SPO, where a circle is divided into two semi-circles along its diameter at a disparity-dependent angle to compute the difference in color distributions between the two semi-circles. The experimental results presented by the authors are among the leading methods in handling occlusions. However, the method's reliance on multiple views poses a challenge, as such an abundance of views may not always be available across all application scenarios, especially camera arrays.

Jeon *et al.* [165] have advanced their optimization-based method presented in [80] by training a combination of different matching cost metrics. Their new method is still focused on specific issues associated with microlens sensors. Therefore, their approach to distortion estimation and correction remains the same and the method is still based on accurate subpixel matching using Fourier Transform. The primary modification is found in the combination of matching costs. To cover a broad range of local scene configurations, they incorporated several consistency measures: SAD calculates the sum of the $\mathcal{L}_1$ loss over a consistent window. In contrast, Zero-mean Normalized Cross Correlation (ZNCC) calculates the cross-correlation after subtracting the mean from a surrounding window. Census Transform (CT) captures local image structures through vectorization, while SAD accentuates higher weights of diverse costs at occlusions. For real world light fields, the most effective combination of these consistency cost functions is unknown. Therefore, a learning framework is employed to determine an ideal combination by fine-tuning the weights. Based on local estimations, a discrete depth label is chosen using random forests: A classifying random forest produces an importance measure for a set consisting of the most credible depth values. Simultaneously, a regression forest calculates a weighted sum across all candidates to guarantee continuous depth prediction. Given the inherent noise in the forest outputs, the estimates are refined using a median filter. For training, the algorithm utilizes 16 scenes from the HCI 4D Light Field Dataset [10]. In order to adapt to light fields recorded with plenoptic cameras, a data augmentation phase is introduced to adapt the input images accordingly. Experimental results show that employing learned matching costs elevates the performance of depth estimation in real images. However, the method is still inaccurate, especially in occluded areas and regions with low texture.

Schilling *et al.* [71] presented a method that leverages the PatchMatch [294] algorithm. For each pixel, one disparity from an initially random set of candidates is chosen according to an objective function. This process is repeated iteratively, pixel by pixel and a new set of disparity candidates is sampled around the chosen values in the neighborhood. Because only one pixel is optimized at a time, favorable candidates can be immediately propagated to adjacent pixels, leading to rapid convergence. The authors model occlusion explicitly to improve the performance at depth discontinuities. One disparity candidate is considered to be occluded by another candidate if a disparity threshold between them is exceeded. The data term is defined as the color variance along the disparity slope of a candidate. The employed smoothness term is based on a modified bilateral filter: Regions that exhibit similar colors and disparities from previous iterations significantly contribute to the smoothness. On the contrary, regions with dissimilar colors or disparities are omitted. This design aids in avoiding oversmoothing at sharp disparity boundaries. The approach demonstrates promising results on both real and synthetic data and even surpasses some deep learning-based methods in performance. However, the method does present some limitations: There is no assurance of achieving global energy minimization due to the reliance on PatchMatch. Additionally, due to the pixel-by-pixel iterative optimization, the runtime is significantly longer compared to deep learning-based methods.

Year	Method	**Light Field**			**Focal Stack**		**Output**		**Design Goals**	
		3D	4D	Plenoptic	Recorded	Rendered	Depth	Refocused	Occlusion	Refocusing
2004	Rajagopalan *et al.* [295]	x			x		x	x		x
2006	Vaish *et al.* [296]		x			x	x		x	
2006	Frese and Gheţa [96]		x		x		x		x	
2007	Gheţa *et al.* [297]		x		x		x		x	
2008	Favaro *et al.* [298]				x		x			
2010	Li *et al.* [299]	x			x		x	x		x
2013	Tao *et al.* [8]			x		x	x		x	
2015	Lin *et al.* [9]			x		x	x		x	

Table 3.2: Classification of defocus-based and combined methods that use EPIs and defocus. Inputs are either three-dimensional, four-dimensional light fields or plenoptic cameras. The focal stack is either directly recorded or rendered from the light field or through synthetic refocusing. Some of the visited methods only predict a depth map, some also do in-focus restoration or synthetic depth of field effect rendering. Defocus-based approaches predominantly target occlusion enhancement, with some extending to synthetic refocusing.

3.1.3 Based on Defocus

Depth from defocus is a technique for estimating depth that extends beyond the exclusive use in light field imaging. It is based on the principle that objects at varying distances from the camera exhibit different levels of focus or blur respectively. Light fields uniquely offer the capability of synthetic refocusing after the initial capture, enabling the application of depth from defocus methods. This section discusses depth-from-defocus approaches that are specifically tailored to light field data.

Vaish *et al.* [296] delve into an analysis of different methods aimed at reconstructing the depth of occluded objects. The primary objective is to enhance robustness in the presence of occlusions. Initially, depth from stereo is compared with depth from focus. To synthetically render refocused images, a large wide-baseline light field is employed. The EPIs are sheared by a number of different disparities. Typically, the energy calculation for stereo correspondences involves computing the variance along the sheared view dimension. However, at occlusions, some of the views capture different points of the occluder, leading to increased variance, even for the correct disparity of the occluded point. This explains why conventional stereo methods often fall short in accurately determining the disparity of occluded points. For synthetic refocusing of an image, EPIs are sheared similarly, by the disparity intended to be in focus. Subsequently, the average color is computed across the view dimension. The amount of high-frequency details in the refocused image can then be used as a cue for the correct disparity. This approach proves more resilient against occlusions; however, the occluder's color still influences the refocused image. Therefore, the authors argue that colors from occluders should be classified as outliers. To achieve this, they introduce two more robust methods: The first approach leverages the median difference between the median color across the view dimension and the actual colors across the view dimension. This way, if more than half of the views contain the color of the occluded point, the energy metric remains close to zero. Their second method employs the Shannon entropy of the color histogram across all view dimensions in the sheared EPI. The underlying rationale is that, for an occluded point, certain views contain a consistent color, which is the color of the occluded point, whereas other views contain varied colors, depicting diverse points on the occluder. Experimental results using a synthetic scene with occlusions accounting for 49% show that the Shannon entropy method excels in reconstructing the occluded depths. However, the authors also note that a combination of different methods could potentially yield even better results.

Favaro *et al.* [298] establish that defocus can be modeled as a diffusion process by using the heat equation. Unlike previous methods, their model of the Point Spread Function (PSF) does not assume the scene to be equifocal, or in other words, all points that are projected to a pixel to be equidistant to the camera. To accommodate this, they introduce a space-varying relative diffusion equation for the PSF. Central to their approach is the concept of *relative blur*. This idea posits that the amount of diffusion that needs to be applied to a more focused image to align with a blurrier one depends on the scene depth. From this premise, they present a depth estimation algorithm: First, depth is initialized as an equifocal plane. Subsequently, for each pixel, a specific amount of diffusion is applied to the sharper image until it matches the blurrier counterpart. The local depth can then be computed from the amount of diffusion applied. For maintaining smoothness in regions lacking texture, they also incorporate an additional smoothness term. Depth optimization is performed iteratively using a gradient descent based algorithm. The method has been validated experimentally on both synthetic and real-world scenes. However, one drawback is that the optimization process is computationally expensive.

Lin *et al.* [9] proposes a method for light field depth estimation solely based on synthetically refocused images. Their method is tailored for plenoptic cameras like Lytro [7] and Raytrix [30]. Depth estimation is based on the local symmetry of the color distribution in a patch around the real depth. While this symmetry assumption holds for smooth areas, it is inaccurate at occlusions. Therefore, a distinction between occluded and non-occluded pixels is required. This is achieved using probabilistic reasoning to produce an occlusion probability map. In the presence of occlusions the cost function is more uncertain, which makes the cost variance a good measure for local occlusion probability. To perform global optimization, the problem is formulated as a Markov Random Field (MRF) with data and smoothness terms. Global optimization is performed using Graph Cut on a resolution pyramid. The authors present promising experimental results on light fields recorded with a Lytro [7] camera. Unfortunately, the refocusing of light fields recorded with plenoptic cameras produces slight artifacts, making the approach not as accurate as most correspondence based methods. Furthermore, the global optimization takes around 20 minutes and estimates only discrete disparities.

3.1.4 Combined

This section focuses on methods that merge EPI analysis with defocus cues for depth estimation. Most of these methods are based on the assumption that defocus-based techniques offer greater robustness, e.g. at occlusions, whereas EPI analysis tends to be more precise. While the superiority of one method over the other can be a subject of debate, numerous publications endeavor to integrate both techniques. The goal of this integration is to combine the strengths of both EPI and defocus approaches within a unified framework.

Schechner and Kiryati [75] conduct a theoretical comparison of DfD with stereo systems. In their analysis, they confirm and refute prevailing theses about the performance of DfD relative to stereo. To draw parallels between the two methods, the authors define an imaging system equipped with a particular focal length and aperture; however, only two pinholes at the lens perimeter permit light to pass. This theoretical system mirrors a stereo system wherein the baseline matches the lens diameter of the lens system. For an in-focus point, projections from both pinholes are projected to a single point on the sensor, resulting in a virtual disparity of zero. Thus, the disparity in a stereo system, corresponds to the size of the blur kernel in the analogous lens system. Using this model, the authors proof that the theoretical sensitivities are equal across both models. Hence, the common perception of stereo systems being more sensitive isn't due to inherent characteristics of both methods, but due to the difference in physical scales. In terms of robustness to occlusions, DfD methods also perform similar to their stereo counterparts. The better performance of DfD in most practical applications, again arises from its smaller scale which minimizes the number of projected points at occlusions. In addition, the authors also demonstrate that DfD, similar to stereo depth estimation, is also based on a matching problem. This is, because contrary to popular belief, an image point doesn't directly map to a single point with the same coordinates in the other image. Lastly, the authors proof the increased robustness of DfD in comparison to stereo. This is attributed to the two-dimensionality of the aperture and therefore contribution of more pixels and more light leading to a superior signal-to-noise ratio. The investigations presented in this article serve as the foundational theory for systems that combine depth from defocus with depth from stereo or light field images.

Rajagopalan *et al.* [295] introduced the concept of merging depth from defocus with stereo depth estimation. To achieve this, two stereo pairs with varying foci are captured, ensuring that both images within a pair maintain similar foci. This allows stereo matching to be applied to each pair and depth from defocus to be applied between the pairs. First, the authors establish the geometric relationship between stereo and defocus. Their method simultaneously restores in-focus images and dense disparity. Therefore, both the defocus and the in-focus images are represented as MRFs. Simulated Annealing (SA) is employed for energy minimization in the process. Their method displays commendable results for depth estimation and in-focus restoration using both real and synthetic images. However, a notable limitation is the necessity to capture a second stereo pair with a different focus, which can be considered a significant downside of the method, especially for non-static scenes.

Frese and Gheţa [96] and Gheţa *et al.* [300] employ a unique approach using a small light field camera array where the foci differ between cameras. Their objective is to fuse stereo and depth from focus techniques. Stereo depth is derived using an energy minimization method based on Graph Cut. The energy function comprises a data term, a smoothness term, and a visibility term to account for occlusions. Depth from focus is estimated based on the assumption that the most in-focus image contains the most high-frequency features. The authors introduce an algorithm to merge stereo depth and focus depth: Initially, depth from stereo is estimated between some image pairs with similar foci via Graph Cut energy minimization. Subsequently, pixels with computed disparities are warped to views where they are expected to be in focus. Depending on the amount of high-frequency features in these views, a low, medium, or high confidence level is assigned to the estimated disparities. In the final step, this confidence is included as an additional energy term and the disparity is optimized in an iterative manner. Experiments show good results on a real-world scene, indicating that the integration of focus cues significantly enhances performance. Nevertheless, their study lacks a comparative analysis with a similar light field where each camera has an identical focus.

In their follow-up publication, Gheţa *et al.* [297] introduce a second method that integrates light field matching with depth from defocus. The energy formulation utilized for stereo matching remains consistent with their previous work. However, their approach for depth from focus is now based on visible color edges. Initially, the Canny edge detector with varied spread parameters is deployed to identify edges, even in defocused areas. Following this, the method models a sharp *virtual* edge and blurs it via a Gaussian point spread function. Both the color and depth attributes of the edge are fitted using Levenberg-Marquardt, to optimize similarity between the *virtual* blurred edge and the image. This derived depth from defocus is subsequently incorporated as an additional term into the stereo energy, which is only non-zero at edges. Experimental results again indicate enhanced performance when focus cues are integrated. Nevertheless, there is still no comparison with a similar light field, where all cameras are set to an identical focus.

Li *et al.* [299] introduced a method termed Dual-Focus Stereo Imaging (DFSI). As an input, two images are captured from different viewpoints, each having distinct foci while maintaining other camera parameters constant. The exact poses and foci are considered unknown, which makes this method usable for freehand snapshots using consumer cameras. The process of defocusing is depicted through a Defocus Kernel Map (DKM), which denotes the per-pixel size of the Gaussian blur disk. A constraint, rooted in the idea that the disparity of in-focus pixels should align with the camera's in-focus disparity, defines the disparity-defocus relationship. To extract both disparity and defocus kernel maps, the algorithm consists of three steps: Initial recovery of camera parameters using only in-focus pixels. Therefore, salient maps from both images are extracted and matched to optimize for the camera extrinsics and foci. In the subsequent phase, the *Defocus Kernel Map Disparity Markov Network* is introduced. While the data and smoothness terms are used similarly to earlier stereo techniques, a direct color comparison between both images is not possible due to the differing foci. As a solution, one image undergoes defocusing based on the disparity difference, ensuring the resultant pair share similar defocusing levels and can therefore be compared accurately. Lastly, the DKM is computed from the disparity and subsequently used to refine the camera parameters. The whole process is repeated two to three times. The practical applications of DFSI are manifold: In the domain of low-light imaging, the wider aperture can be used to capture more light. The shallow depth of field is then counteracted by combining the in-focus regions of both images. Automatic defocus matting facilitates foreground-background differentiation, ideal for placing the foreground object onto a new background. To achieve this, a trimap is derived from the DKM. For multifocus photomontage, the method is used to fuse in-focus regions. This, again, produces an image with a deep depth of field, synthesized from shallow depth of field inputs. The authors presented good results on some real photographs. However, the method struggles near occlusions, because the point spread function in those regions is significantly different from the modeled Gaussian blur disk.

Tao *et al.* [8] present another method that combines defocus and correspondence search, tailored specifically for plenoptic light field cameras. They propose a combination of defocus cues and correspondence cues using a straightforward shearing approach: The EPIs undergo gradual shearing by minor disparity steps and are analyzed in two distinct ways: Firstly, in the defocus analysis, the EPI is integrated along the view dimension. Subsequently, the gradient along the image dimension is computed. This procedure effectively refocuses the light field at a particular disparity and then calculates the amount of high-frequency details. A large amount of high-frequency details indicates the true disparity. Secondly, in the correspondence analysis, the variance along the view dimension of the sheared EPI is computed. Here, a small variance acts as a cue for the correct disparity. The authors emphasize that defocus and correspondence cues each have unique advantages and shortcomings. Consequently, their fusion offers a synergistic effect that improves the results. Specifically, defocus is commonly known to be more robust against occlusions, repeating patterns, and noise, while correspondences display enhanced performance with robust high-contrast textures. Because this is one of the first publications that focus on depth estimation from Lytro [7] cameras, the authors further introduce two additional applications: synthetic depth-of-field effects and matting. Their experiments show relatively good results on synthetic and real light fields. However, the approach of fusing defocus and correspondence cues to combine the advantages of both methods remains questionable. As proven by Schechner and Kiryati [75], when utilizing consistent physical camera dimensions, both techniques should have comparable characteristics. In addition, due to the small baseline of the Lytro [7] camera, extreme depths cannot be reconstructed, using the proposed method.

3.1.5 Based on Deep Neural Networks

Year	**Method**	**Input**					**Design Goals**			
		View	Cross	Star	Full	Focal Stack	Unsupervised	Occlusion	Accuracy	Runtime
2016	Heber and Pock [89]		x							
2016	Heber *et al.* [90]		x							
2017	Heber *et al.* [134]		x							x
2018	Shin *et al.* [20]			x					x	x
2019	Zhou *et al.* [92]					x		x		
2019	Zhou *et al.* [93]			x		x		x		
2019	Zhou *et al.* [206]	x					x			
2020	Tsai *et al.* [21]				x				x	
2021	Li *et al.* [56]		x							x
2021	Chen *et al.* [91]			x				x		
2023	Chen *et al.* [272]				x				x	

Table 3.3: Classification of neural network based methods. Inputs to these networks are often subsets of the light field: either a single view, a horizontal and vertical stack (cross configuration), two additional diagonal stacks (star configuration), the full light field or a synthesized focal stack. Objectives of the explored methods include enabling unsupervised training, improving occlusions, improving overall disparity accuracy or shortening the runtime.

This section highlights the most recent advancements in the field of light field depth estimation, focusing on methods based on deep learning. These deep learning approaches do not depend on manually crafted objective functions, but instead, they learn EPI analysis, focal stack analysis, and smoothness directly from the data. Compared to classical methods, most deep learning-based approaches also offer significantly faster runtime. This efficiency makes them more suitable for a broader range of applications.

In 2016, Heber and Pock [89] presented the pioneering deep learning approach for depth estimation from four-dimensional light fields. Their method comprises two primary steps: Initially, a neural network is used to predict a pixel's disparity based on a horizontal and a vertical EPI patch. Subsequently, these disparities undergo refinement through global optimization. The input to their neural network architecture are two EPI patches, one horizontal, one vertical, centered around the target pixel. Following four convolutional layers and one fully-connected layer, the output is a single scalar value representing the estimated disparity for this pixel. Because no large light field datasets with ground truth depth existed at the time, the authors also introduced a randomly generated dataset. All scenes in this dataset contain randomly arranged objects, heavily occluding each others, on random backgrounds. For rendering the scenes, the raytracing software *POV-Ray* was used. To enhance generalization, the authors also employed data augmentation strategies: random changes in hue and luminance, and adding noise. Given that the network only predicts a singular disparity for each pixel individually, the predicted disparity maps tend to be very noisy. Therefore an additional refinement model is introduced. The network outputs serve as the data term and an additional smoothness term is added. To globally optimize the disparity map, the primal-dual algorithm is used. Experimental results demonstrate that the refined outputs outperform other methods at the time. However, one notable limitation lies in the neural network's need to infer each pixel independently. Coupled with the refinement optimization, this results in a significantly longer runtime, especially when compared with later deep learning-based methods.

To address the aforementioned limitations, the same authors introduced a follow-up approach [90] that leverages the U-Net [301] architecture. This refined approach predicts the disparity of entire EPIs collectively, rather than on a pixel-by-pixel basis. Structurally, the architecture consists of an encoder and a decoder subnetwork, interconnected by residual pinhole connections. For training, the authors employed the identical randomly generated synthetic dataset. Experiments show visual and quantitative enhancements. However, because each line or column of the light field is predicted separately, the disparity maps suffer from streaking artifacts along the image dimensions.

In their next follow-up publication, Heber *et al.* [134] addressed these streaking artifacts. Instead of predicting the disparity for a single EPI along one line or column of the light field, the revised network architecture estimates disparities across an entire EPI volume. Therefore, all views are stacked along a single view dimension to form an EPI volume that serves as input to the network. To enable operations on EPI volumes instead of individual EPIs, the network adopts three-dimensional convolutions instead of two-dimensional convolutions. Despite these changes, the overall design of the architecture remains heavily influenced by U-Net [301]. The network outputs a disparity estimation for each pixel within the three-dimensional EPI volume. Outputs for each volume, for instance, for the horizontal and vertical volumes in a cross-configuration light field array, are inferred independently and subsequently combined to a final disparity map. Experimental results indicate that this method not only eradicates the streaking artifacts seen in their preceding work [90], but also offers competitive performance in relation to other methods at the time. An added advantage is the reduced inference time, compared to the previous method that inferred each EPI separately.

Shin *et al.* [20] proposed *EPINET*, a method for light field depth estimation utilizing a fully-convolutional architecture. The authors propose a multi-stream backbone network tailored to separately infer features from horizontal, vertical, and two diagonal light field stacks. They do not leverage the entire four-dimensional light field, instead opting for a star-shaped subset. Subsequent to processing, all outputs from these individual streams are concatenated, serving as input to the neural network head. This head is structured around eight identical blocks: each of these comprises two convolutional layers, succeeded by a Rectified Linear Unit (ReLU) activation and a BatchNorm layer. Concluding the architecture, the last layer of the network is designed to output the per-pixel disparity of the center view. For training, the HCI 4D Light Field Dataset [10] is employed. Due to the dataset's limited number of scenes, extensive augmentation became necessary. One noteworthy augmentation method involves adjusting the scene orientation by incorporating only 7×7 views from the available 9×9 views. This way, using the extra views, the light field can be repositioned. Beyond this, they also implement more conventional augmentation techniques, including rotation, flipping, image and color scaling, as well as random adjustments to contrast and brightness. The proposed method demonstrated state-of-the-art performance on the HCI 4D Light Field Benchmark [10]. Yet, the proposed architecture shows an important limitation: the network's receptive field effectively restricts the maximum disparity it can predict. This becomes a significant hindrance for wide-baseline high-resolution camera arrays, which often exhibit disparity ranges of hundreds of pixels.

Zhou *et al.* [92] introduce an approach to estimate disparity from a focal stack using a neural network. Rather than regressing continuous disparities, they opt for pixel-wise classification of discrete disparity labels. For each disparity label, an image refocused to this disparity is synthetically rendered. Their network architecture is split into two sub-networks. The first sub-network employs three-dimensional convolutions to extract features from the whole focal stack, premised on the assumption that the stack is symmetric around the true disparity. Concurrently, the second sub-network operates exclusively on the central view to draw structure information from the two-dimensional image. Such extraction from the central view is beneficial, given that unlike the focal stack elements, all depths are in focus simultaneously. Features extracted from both sub-networks are concatenated and two fully connected layers predict logits for each discrete disparity label. The network undergoes training and evaluation using the HCI 4D Light Field Dataset [10], in addition to several real recordings. Experimental results show that the performance is comparable to EPINET [20].

In their follow-up publication, Zhou *et al.* [93] introduce a network architecture that aims to combine depth cues from a focal stack with depth cues from EPIs. This is grounded in the assumption that correspondence and defocus cues are complementary, as supported by prior research. Again, they employ pixel-wise classification of discrete disparity labels, instead of continuous regression. The proposed architecture consists of three sub-networks: First, a *defocus pathway*, similar to the focal stack sub-network from their previous publication [92]. Next, the *structure pathway*, designed to extract structural features from the center view. Lastly, the *EPIs pathway*, to extract features from four EPI volumes: horizontal, vertical, and two diagonal orientations, similar to EPINET [20]. Initially, cues from the defocus and structure pathways are merged using a *feature-level fusion* head. Subsequently, these results are again merged with the EPI cues through a *prediction level fusion*. Lastly, the network outputs the per-pixel probability of each disparity label. The architecture is trained and evaluated using the HCI 4D Light Field Dataset [10] and some real recordings, showing, again, a performance on par with EPINET [20]. However, a constraint of this model is, again, its restriction to narrow-baseline light fields. One could also argue that, given the shift-operation introduced in chapter 4 of this thesis, the network might autonomously learn to use the three cues, focus, structure, and EPI correspondences, without the need for a specifically tailored architecture. Moreover, the complexity of the proposed structure arguably makes it computationally heavy and presents a huge challenge for hyperparameter tuning.

In their second follow-up publication, Zhou *et al.* [206] introduced one of the first unsupervised depth estimation technique based on light fields. Their method employs an encoder-decoder architecture, which predicts a disparity map using solely the center view. Therefore, technically speaking, their approach aligns more with monocular depth estimation. For unsupervised training, they utilize three distinctive loss functions: First, a photometric loss is calculated between views warped to other views by the predicted disparity and the true views. Second, a defocus loss is incorporated by generating a synthetic all-focused image via warping all views, averaging their colors, and comparing this synthetically rendered all-focused image with the center view. Lastly, a symmetry loss is applied, relying on the disparity map predicted for the center view and the adjacent views. Experimental results indicate inferior quantitative results compared to supervised and most optimization-based approaches. Moreover, there's a contention that this method is essentially a monocular depth estimation technique, albeit trained on light fields in an unsupervised manner. It's noteworthy that the authors confirmed using all 28 scenes from the HCI 4D Light Field Benchmark [10] for training. This raises concerns, as the evaluation on the very same benchmark scenes becomes questionable.

Tsai *et al.* [21] address a limitation in previous deep learning techniques where only a subset of the four-dimensional light field was used. Given that a four-dimensional light field encompasses a high degree of redundancy, pinpointing the views that provide valuable cues for depth estimation becomes challenging. Therefore, the authors propose to train an attention map, guiding the contribution from each view to the final result. The neural network architecture they introduce is structured into three distinct parts: Initially, features are isolated from each view independently. This extraction employs a sequence of convolutions coupled with a Spatial Pyramid Pooling (SPP) module. The SPP module effectively increases the receptive field, enabling even large non-textured regions to contain meaningful features propagated from neighboring textured zones. Subsequently, the features extracted from all views undergo a shift by a designated number of discrete disparity steps and are concatenated into a cost volume. From this cost volume, an *attention module* is utilized to predict the attention map. The features extracted from each view undergo a weighting based on their respective attentions. Lastly, a set of convolutions are applied for disparity estimation from the weighted cost volume. A softmax operation conducts a weighted sum across all discrete disparity labels, to regress a continuous disparity value for each pixel of the center view. Upon evaluation, experiments conducted on two distinct datasets indicate that the architecture can outperform EPINET [20] in terms of accuracy while having a slightly longer runtime. Further, an ablation study confirms that imposing constraints on the attention map, such as maintaining a certain symmetry yields optimal outcomes. However, the authors also highlight a potential drawback: the architecture exhibits weaknesses in specular and textureless areas.

Li *et al.* [56] introduce a light-weight network tailored to process wide-baseline light fields. A cornerstone of their approach is the drastic reduction of parameters, bringing them down to approximately 1.8 million, coupled with the construction of a cost volume derived from feature maps. The network architecture starts with extracting a feature map from each distinct view. Following this, the extracted feature maps undergo a shift by a set number of discrete disparity steps. A cost volume is then constructed using the absolute difference between the shifted features and the features of the center view. This computed cost is then aggregated over all the discrete disparities. To further refine the results, an additional attention block is deployed to assign weights to the costs originating from different views. This attention mechanism, similar to prior research, is grounded in the observation that certain directions of a light field, such as horizontal views, might present occlusions for a certain pixel. For the same pixel, other directions, like vertical views, align parallel to the occlusion boundary and therefore bypass the occluder. Lastly, a continuous disparity is obtained using soft argmin. To train their network on wide-baseline light fields, the authors also created a synthetic wide-baseline dataset. This dataset encompasses 36 hand-crafted and 345 randomly generated scenes. Experimental evaluations underscore that the proposed method performs well on wide-baseline, but also narrow-baseline light fields. Inference time is also very low due to the small number of parameters.

Chen *et al.* [91] extend the concept of view attention based on ideas previously presented by Tsai *et al.* [21] and Zhang *et al.* [72]. Their contribution is specifically motivated by occlusions, especially when the occluding object possesses a relatively straight edge. In that case, in certain orientations of the light field, e.g. the horizontal views, the occluder might be included, undermining the reliability of disparity prediction. Conversely, other orientations, like the vertical views, might exclude the occluder, making predictions more reliable. To exploit this, the authors propose an approach to first extract features along four distinct orientations: horizontal, vertical, and two diagonals, referred to as the star configuration. The cost volumes of these orientations are then merged in a manner akin to Tsai *et al.* [21], based on a dynamically predicted attention. However, in contrast to [21], this merging isn't global; attention is predicted on a per-pixel basis, allowing for adaptability to local occlusions. Moreover, attention is also applied within each orientation individually. This internal attention is motivated similarly: if an occlusion is visible in the views on one side of an orientation, these views receive less attention. The proposed neural network architecture can be divided into four parts: First, features are extracted and cost volumes are constructed from each of the four orientations, with the inclusion of a shift operation, similar to the technique in Tsai *et al.* [21], to form these volumes. Second, features of each orientation are merged where views to the left, center and right can receive different attention levels. In the third step, features from all four orientations are merged similarly. Lastly, cost is aggregated, resulting in a pixel-wise disparity map as the final output. Their experiments, applied to both synthetic and real-world datasets, showcase state-of-the-art performance.

Chen *et al.* [272] focus on sub-pixel disparity refinement. Therefore, they introduce the BadPix Correction and post-refinement Network (BpCNet). At its core, BpCNet's goal is to refine the disparity within a narrowly defined window of an initial disparity map. This initial disparity map is assumed to be pre-determined, e.g. by another disparity estimation method. For data augmentation, several techniques are employed: First, rotation, contrast adjustment, brightness modulation, and color re-distribution. Furthermore, noise is added to the initial disparity estimates. To augment their initial disparity input even more they use an iterative training process, whereby the refined output disparity is utilized as the initialization for subsequent training iterations. For disparity refinement, a predefined number of hypotheses is formed around the initial disparity. Following this, a dedicated feature extraction network retrieves features from every light field view. Based on these feature maps, a cost volume is constructed. One point of contention raised by the authors is the inaccuracy of bilinear or bicubic interpolation for handling sub-pixel shifts. To address this, they draw inspiration from Jeon *et al.* [80] and apply a phase-shift based interpolation. For efficiency's sake, they limit the interpolation to a small window around each pixel. Inspired by previous methods, they utilize an attention decoder to extract an attention map from the derived cost volume. The final refined disparity map is then obtained, using weighted fusion that integrates the attention maps. Empirical tests, conducted on the HCI 4D Light Field Dataset [10] and real-world captures demonstrate BpCNet's ability to successfully refine the outputs of existing disparity estimation algorithms.

3.1.6 Datasets and Benchmarks

The field of light field depth estimation experienced a surge in research activity, as illustrated in fig. 3.1, leading to the introduction of various datasets tailored to specific tasks. Particularly with the rise of deep learning, the community has shifted towards larger, synthetic, and often randomly generated datasets. As interest in this field of research grows, benchmarks have become essential for evaluating and comparing the ever-increasing number of methods.

One of the first datasets is *The (old) Stanford Light Field Archive*, first used by Levoy and Hanrahan [3] in their foundational work on light field rendering. This dataset comprises four synthetic scenes at diverse resolutions and includes two real scenes captured with a video camera and a computer-controlled gantry.

Another notable contribution is from Joshi [4], who published a three-dimensional light field dataset using two setups: one with a linear array of eight Video Graphics Array (VGA) cameras, and another with a computer-controlled gantry holding a DSLR camera on a linear translation stage. This dataset includes nine video light fields and seven static light fields.

The (new) Stanford Light Field Archive [5] further expanded the field with recordings from four different sources: the Stanford Multi-Camera Array, composed of 100 VGA video cameras; the light field gantry, initially built for the *Digital Michelangelo Project*, later adapted to record a light field of *Michelangelo's statue of Night*; an inexpensive gantry based on *Lego Mindstorms*, surprisingly accurate;

and the Stanford Light Field Microscopy project, which implemented a microlens array in a conventional optical microscope for microscopic light field recordings.

Significant synthetic light field datasets were also introduced by Wetzstein [55] and Marwah *et al.* [105], offering 14 synthetically generated light fields, some tailored for three-dimensional displays and others simulating light field cameras, rendered through *POV-Ray* raytracing.

Wanner *et al.* [6] launched the first public light field depth estimation benchmark, motivated by the availability of consumer plenoptic cameras but the scarcity of narrow-baseline datasets that match these cameras' characteristics. Their dataset included seven synthetic scenes with ground truth and six real-world light fields, captured using a gantry, with ground truth acquired via a structured light scanner. They also provided a Compute Unified Device Architecture (CUDA) C library, enabling comparisons with newly published algorithms.

The Stanford Lytro Light Field Archive [302], with its hundreds of light fields recorded using the *Lytro Illum* [7] camera, covers various categories like *Flowers & Plants*, *Cars*, and more challenging examples in *Occlusions* and *Refractive & Reflective surfaces.* This dataset also includes depth maps estimated by Lytro's commercial software.

Heber and Pock [89] contributed a unique, randomly generated synthetic dataset, comprising 25 scenes with objects from a selection of 20, rendered via *POV-Ray.*

The most widely adopted benchmark for light field depth estimation was introduced by Honauer *et al.* [10], featuring 24 handcrafted scenes. These include four stratified scenes addressing specific challenges, four test scenes, four training scenes, and twelve additional scenes not part of the benchmark. Rendered using Blender's *Cycles* renderer, this benchmark introduced novel error metrics and evaluations, a benchmarking website, and an evaluation toolkit. It also provided an initial performance analysis of four state-of-the-art algorithms. However, the advent of deep learning-based methods necessitated larger training datasets, prompting some publications to introduce their own larger datasets.

Feng *et al.* [11] supplied a real dataset with ground truth, scanning 19 objects using the *3dMD* scanner, then placing them in outdoor scenes recorded with a Lytro Illum [7] camera and rendering the depth to obtain ground truth depth maps.

The Stanford Multiview Light Field Datasets by Dansereau *et al.* [184] included a dataset captured with the Lytro Illum [7] camera, featuring 4211 light fields across 30 categories, and a *Three-View Dataset* captured with three Lytro Illum Cameras.

Li *et al.* [56] created another synthetic dataset for deep-learning applications, consisting of 36 handcrafted and 345 randomly generated scenes with flying objects, rendered using Blender's *Cycles* renderer.

The urban light field dataset by Sheng *et al.* [12], focusing on semantic segmentation tasks, consists of 1074 scenes: 824 real-world scenes recorded with a Lytro Illum [7] camera and 250 synthetic scenes modeled after the real scenes and rendered using Blender's *Cycles* renderer. This dataset includes semantic segmentation annotations, but only the synthetic scenes have ground truth depth.

Lastly, Sheng *et al.* [282] organized the LFNAT 2023 challenge, a significant event in the field with 75 participants submitting methods for light field depth estimation based on the synthetic UrbanLF dataset [12]. The challenge highlighted the continuous evolution and expanding breadth of research in this dynamic field. They published the ranking of the top seven finalist teams.

3.2 Light Field Rendering

Another task closely related to depth estimation from light fields is light field rendering. Light field rendering displays the light field from a new view point, often using depth information to enhance image quality. It plays a crucial role in applications such as VR, AR, and advanced photography. The methods and technologies in light field rendering have evolved alongside depth estimation techniques, benefiting from the increased understanding and precision in depth information. This section provides a concise overview of the key milestones in this field.

Adelson and Bergen [23] formalize the concept of the plenoptic function in their work. This function embodies all visual information that is observable. It models light as rays passing through any point in space, a concept already addressed by Leonardo da Vinci, who termed it the *radiant pyramid*. The plenoptic function encompasses all possible view positions and directions, spanning every moment and across all wavelengths. They defined the plenoptic function which includes orientation, light wavelength, time, and position. Since its introduction, the plenoptic function has been extensively utilized in computer vision and adapted to suit various specific applications. In the context of light field depth estimation, a four-dimensional definition focusing on position and orientation is commonly employed, often omitting time and color, or alternatively, modeling red, green, and blue color information separately.

In their exploration of the plenoptic function, McMillan and Bishop [303] introduce a novel framework for Image-Based Rendering (IBR). IBR relies primarily on photographic data instead of detailed geometric models. As a result, this method significantly reduces computational complexity, because it only manipulates two-dimensional images through warping and interpolating, rather than processing an extensive number of geometric primitives. However, as a consequence, it also offers less flexibility in deviating from real scenes. The contributed IBR method utilizes cylindrical projection for sample representation. This choice is made due to the storage efficiency and the ease it provides in image analysis, despite its inherent limitations in capturing the vertical field of view. The process uses manual annotations on a set of points in two neighboring images to establish geometric constraints in the form of image flow fields. To render a novel image, the plenoptic function for the desired viewpoint is reconstructed. This reconstruction process models occlusion and interpolates between reference images, offering a new perspective derived from existing data. The method is demonstrated on two scenes recorded using a video camera and a panning tripod.

Levoy and Hanrahan [3] present an IBR technique simply called *Light Field Rendering*. They introduce a two-plane representation, known as the light slab representation. This representation is particularly well-suited for light fields captured using a gantry or light field array. Due to the still sparse sampling of these light fields, aliasing can be an issue. To mitigate this, the authors introduce a virtual aperture, the size of which is equal to the uv sample spacing. Addressing the memory limitations of the era, a light field compression algorithm is introduced, achieving an approximate compression rate of $100 : 1$. The paper also introduces a real-time image viewer, within which different interpolation methods are evaluated. Their results are demonstrated on multiple re-rendered synthetic light fields, showcasing the practical applications of their approach.

Gortler *et al.* [304] introduce *The Lumigraph*, building upon concepts similar to those found in [3]. They also utilize the two-plane representation to parameterize the plenoptic function within a finite empty space cube. To translate a set of recorded images into the Lumigraph, they propose a discrete version of this plenoptic function, assigning values to a discrete number of points. For the recovery of a continuous value for the Lumigraph from these discrete values, they define a basis function that samples from the discrete plenoptic function. They also present the notion that the quality of IBR can be enhanced with some knowledge of the scene's three-dimensional structure. Accordingly, they modify the basis function for scenes with known geometry. Additionally, they describe a system for capturing a Lumigraph, which involves a rough approximation of the three-dimensional shape based on the segmentation of an object's silhouette. Furthermore, they introduce a technique for efficiently rendering their Lumigraph representation using texture mapping hardware. Their method is demonstrated across a variety of real and synthetic scenes.

Over the last 25 years, a vast number of IBR techniques has emerged, building on the fundamental ideas in the field. Chai *et al.* [305] determined the minimum sampling rate of light field signals, taking the depth of the scene into account. Geys *et al.* [306] made a significant contribution with a fast, Graphics Processing Unit (GPU) based algorithm for depth estimation, leading to more efficient rendering. Siu and Lau [307] introduced a novel method for image registration in IBR, utilizing triangular patches. Kubota *et al.* [308] presented an approach that bypasses the need for explicit depth information, instead filtering all-in-focus intermediate images between cameras. Georgiev and Lumsdaine [309] focused on the challenges of artifacts in IBR specifically caused by focused plenoptic cameras. Bishop and Favaro [100] enhanced the rendering results of images captured with plenoptic cameras by employing an explicit image formation model. Mildenhall *et al.* [74] innovated by developing a system for capturing a sparse light field of a static scene using just a smartphone. Building on this, they introduced the concept of MPIs, which are predicted using a neural network. This allows for rendering the light field at a higher sampling rate than was previously possible. In a groundbreaking contribution, Mildenhall *et al.* [27] employed a neural network to learn the plenoptic function directly, naming their representation a *Neural Radiance Field (NeRF)*. This work not only achieved outstanding visual results but also laid the foundation for a new line of research.

3.3 Posterior Regression

Traditional depth estimation methods predict only a single depth value per pixel. These methods assume that each pixel only receives light from one point in the scene. While this assumption holds if the scene only contains Lambertian and opaque objects, it does not in the presence of specular or semi-transparent surfaces. Therefore, this section focuses on works that estimate a whole posterior for each pixel, in depth estimation and other fields. By estimating a whole posterior, these methods can better account for uncertainties and ambiguities inherent in real-world data. The ability to estimate a whole posterior rather than a single depth value per pixel enhances the robustness and reliability of depth estimation in dynamic and unpredictable environments and is useful for many downstream applications.

The concept of Bayesian Neural Networks (BNNs), as introduced by MacKay [310], represents a significant shift in how neural networks are understood and implemented. In traditional neural network frameworks, the focus is on predicting a single, specific output based on the given input. However, BNNs introduce a probabilistic interpretation to this process. In a BNN, instead of aiming for a singular deterministic output, the network models the likelihood of various possible outputs given the input. This is achieved by considering the posterior probability of the network weights, conditioned on the dataset. The network learns not just to predict an output, but to estimate a distribution of possible outputs, providing a measure of uncertainty in its predictions. The process involves applying Bayes' rule, combining the likelihood maximized during training with a prior distribution over the network weights. This approach offers a more nuanced understanding of the model's predictions, accounting for the uncertainty inherent in real-world data and the learning process. However, one of the challenges with BNNs, particularly in large-scale networks, is computational efficiency. As the size and complexity of the network increase, efficiently computing the posterior distribution of the weights becomes increasingly difficult. This limitation has been a significant hurdle in the wider adoption of Bayesian methods in large-scale neural network applications. Despite these challenges, the Bayesian approach provides a powerful framework for dealing with uncertainty and incorporating prior knowledge into neural network models.

Neal [311] introduced the Markov Chain Monte Carlo (MCMC) method to approximate the posterior over the model weights. He also analyzed the significance of the chosen prior when dealing with a high number of model weights. These methods facilitated an efficient approximation, enabling the training of BNNs. However, it is important to note that even with these advancements, MCMC methods have limitations. Even recent MCMC approaches are still restricted by their applicability to a limited number of dimensions. This constraint highlights a critical challenge in scaling BNNs for more complex problems.

Gal and Ghahramani [312] approximate a Gaussian Process by utilizing dropout. This approach aims to capture epistemic uncertainty, proposing a novel approximation of the marginalization over all possible networks. They demonstrate that conducting several forward passes with dropout enabled can approximate a BNN. This approximation is achieved using Monte-Carlo integration. In this process, dropout during a forward pass is interpreted as a sample from the posterior distribution of all networks. Unlike predictive models with an explicit variance output to quantify uncertainty, this Monte-Carlo-Dropout technique also captures uncertainty inherent in the model itself. This inherent uncertainty is primarily caused by a lack of training data. The technique provides a more comprehensive uncertainty estimation by considering both the data and the model.

Kendall and Gal [313] analyzed different types of uncertainty relevant for Computer Vision. They propose capturing aleatoric uncertainty, uncertainty inherent in the data, by training the network to predict a variance for its output. This approach involves modeling a normal distribution over the outputs, which can also be interpreted as learned loss attenuation. Epistemic uncertainty, uncertainty inherent in the model, which diminishes with an infinite amount of data, is inferred using the Monte-Carlo Dropout technique as described by Gal and Ghahramani [312]. The authors developed a single model capable of inferring both types of uncertainties. They conclude that these uncertainties are helpful, but not mutually exclusive.

Ilg *et al.* [314] conduct a comparative analysis of various uncertainty quantification methods in optical flow, a technique used for tracking motion in video sequences. They compare the effectiveness of aleatoric uncertainty quantification in a single network against that in an ensemble of networks, and also against a single network with multiple output heads trained using the Winner-Takes-It-All loss. Their study reveals that Monte-Carlo Dropout, a technique commonly used for measuring uncertainty, exhibits limited performance in regression tasks. Additionally, they find that ensembling does not significantly enhance uncertainty quantification compared to using a single network. However, one could argue that this outcome may be attributed to the minimal epistemic uncertainty in optical flow tasks, particularly when employing large synthetic datasets. As Kendall and Gal [313] has illustrated, methods like Monte Carlo Dropout capture epistemic uncertainty, which diminishes with the availability of extensive data. Given the substantial size of the synthetic datasets used in their research, the epistemic uncertainty is likely to be negligible. Consequently, there is limited scope for these methods to outperform a single network in such contexts. This analysis implies that the performance of different uncertainty quantification methods is highly dependent on the specific nature of the data and the task at hand. In scenarios involving large datasets, where model uncertainty is inherently low, the advantage of certain methods may not be as apparent.

Ardizzone *et al.* [315] introduced the use of Invertible Neural Networks (INNs) for estimating the posterior distribution over the predictions made by a neural network. They accomplished this by training the inverse function, which effectively maps desired outputs back to the inputs and a latent space. This latent space is designed to capture all information not included in the input. By ensuring that the latent space adheres to a multivariate Gaussian distribution, the posterior distribution can be approximated. This approximation is done by sampling from the latent space during the forward pass of the network. Using multiple sampled forward passes, the posterior distribution can then be accurately fitted.

In their follow-up work, Ardizzone *et al.* [316] enhance the utility of INNs by incorporating the Information Bottleneck (IB) principle into their training regime, particularly for the task of image classification. To achieve this, they map the input image to a Gaussian Mixture Model (GMM). A GMM is a probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions, each representing a different group or class. In the context of their work, each component of this mixture model represents a different class label. When an image is fed into the network, the INN assigns probabilities to each of these Gaussian distributions, essentially determining how likely it is that the image belongs to each class. This method of classification is more nuanced than traditional approaches because it accounts for the probabilities of an image belonging to multiple classes simultaneously, instead of forcing a single class prediction. Using Bayes Rule, the exact posterior distribution over all class labels can be derived. This approach provides a more comprehensive and probabilistic understanding of the classification task, capturing the uncertainties and complexities inherent in real-world image classification. The advantage of this method is its efficiency, requiring only a single forward pass through the network. This is a significant improvement over many Bayesian deep learning frameworks, which often require multiple passes or more complex computations to estimate uncertainties or probabilities.

4. Wide-Baseline Light Field Depth Estimation with EPI-Shift

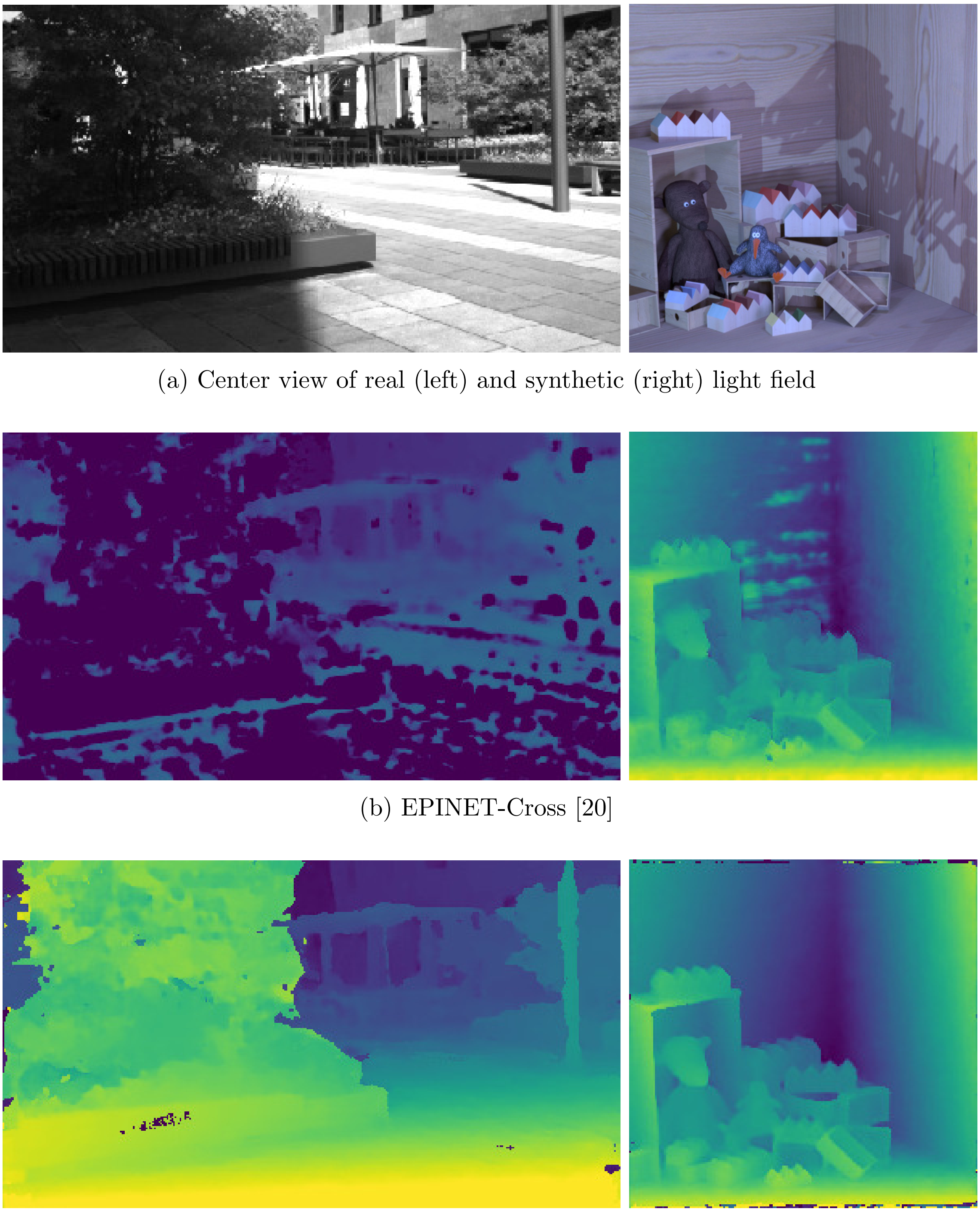
This chapter introduces *EPI-Shift*, a method for applying deep learning-based light field depth estimation to wide baselines. EPI-Shift expands the scope of application to larger light field cameras and therefore enables new uses in many fields.

4.1 Introduction

There are many types of light field cameras, as shown in section 2.2. One common type are plenoptic cameras based on a microlens array [7] which have a rather limited resolution and baseline. Multi-camera light field arrays are more expensive, bulkier, harder to use and to calibrate. However, once these challenges are mastered, the major advantage of these systems is their accuracy, which grows linearly with the baseline between the cameras. Therefore, most camera arrays allow for a much higher reconstruction accuracy compared to the more compact plenoptic cameras. As this is useful for many applications, like the acquisition of training data for other depth estimation methods, this chapter focuses specifically on wide-baseline light field depth estimation.

We propose a learning-based light field depth estimation method. This is challenging due to the non-availability of real-world training data. The creation of real-world reference depth is problematic, as no other dense measurement approach is more accurate than light field depth estimation. For example, structured light scanning is problematic in the context of occlusions and LIDAR is considerably more sparse than light field data. Therefore, all training data is synthetic and the pool of publicly available datasets is small. Also, all these training images have a small baseline, emulating a micro-lens based camera rather than a camera array. Hence, current learning-based approaches fail dramatically for wide-baseline light fields, as exemplified in the real world scene in fig. 4.1. Interestingly, even for smaller disparities, the performance is limited by poor generalization, as demonstrated in the synthetic scene in fig. 4.1, where artifacts appear *within* the trained disparity range.

One major cause for this problem, the limited receptive field of previous methods, is illustrated in fig. 4.2. Expanding the receptive field would cause worse generalization performance. However, by applying EPI-Shifts to the input of our neural network, we circumvent this flaw. The basic idea is inspired by the technique of plane-sweep volumes. Instead of directly estimating the disparity from the light field image stacks, we utilize a plane, sweeping through space, as common for stereo and multi-view depth estimation [73], [317]. Hence, we split the task into classification and regression.



(a) Center view of real (left) and synthetic (right) light field

(b) EPINET-Cross [20]

(c) Our EPI-Shift

Figure 4.1: Light Field Depth Estimation. (a) A *real* light field (left) with a large disparity range of $[0, 12]$ and a *synthetic* light field (right) with a small disparity range of $[-2, 2]$. (b) The current state-of-the-art method EPINET-Cross [20] has only been trained for the small disparity range. It therefore fails at extreme disparities in the synthetic image (right, background) and outside of the trained range in the real image (left, foreground). (c) Our EPI-Shift approach performs well for both, small and large disparities. Due to better generalization, it even outperforms EPINET-Cross for small disparity ranges.

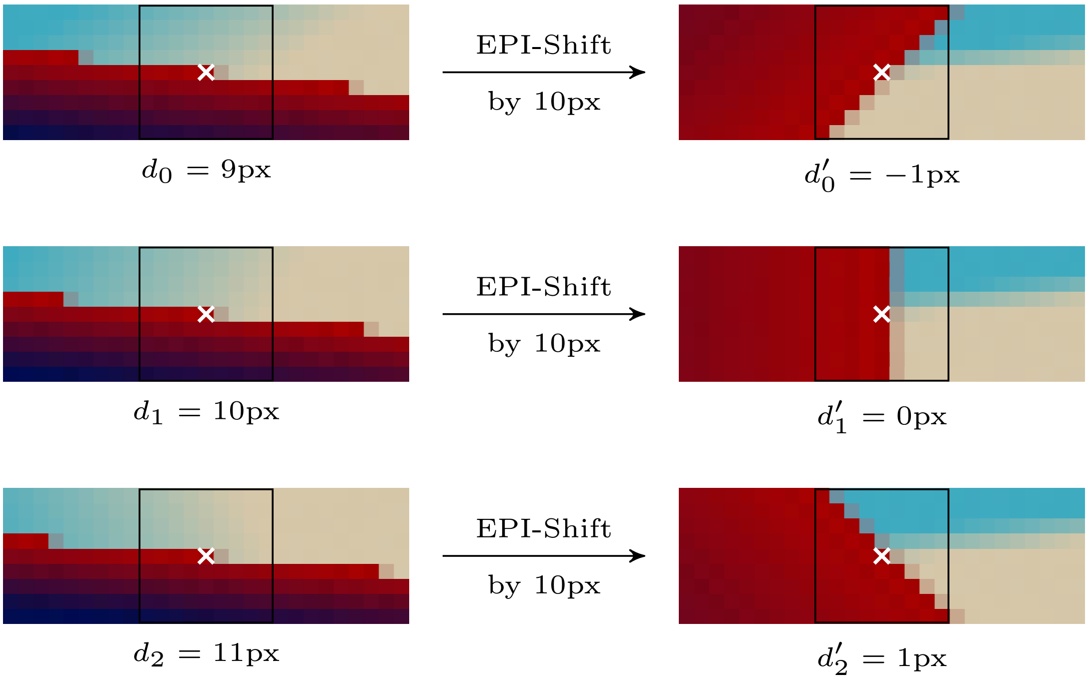


Figure 4.2: Idea of EPI-Shift. Three EPIs (left), consisting of horizontal lines from different views. The CNN's task is to estimate the correct disparity for the center pixel (white cross) by predicting the slope of the line going through this pixel. Note the nearly invisible difference between a disparity of 9, 10 or 11 pixels which can only be estimated using a large receptive field. After applying an EPI-Shift of 10 pixels (right), the difference is clearly visible. Therefore, our network only requires a minimal receptive field, visualized as rectangular box, to classify whether the shifted disparity lies within ±1 pixels and regress a sub-pixel accurate disparity offset.

The classification map states, per pixel, whether objects observed at tested plane sweep are within a refocused disparity range of $[-0.5, 0.5]$. It is used to merge all independent estimates from the plane sweep volume, while the disparity regression provides sub-pixel accuracy. This approach considerably improves generalization, as we are now able to infer the depth of a wide-baseline scene with a network, trained solely with small-baseline training data.

Let us summarize our main contributions:

- Applying the idea of plane-sweep volumes in the context of light fields, which we denote as EPI-Shift. This enables learning-based approaches to generalize well to large-disparity test data, even with small-disparity training data.

- A network architecture, which enables improved long-range reasoning by combining a feature extraction network [20] with a subsequent U-Net architecture [90], [301] for excellent long-range smoothing with low artifacts.

- Our approach outperforms the state-of-the-art learning-based approaches with the same input modality and is on par with hand-crafted methods.

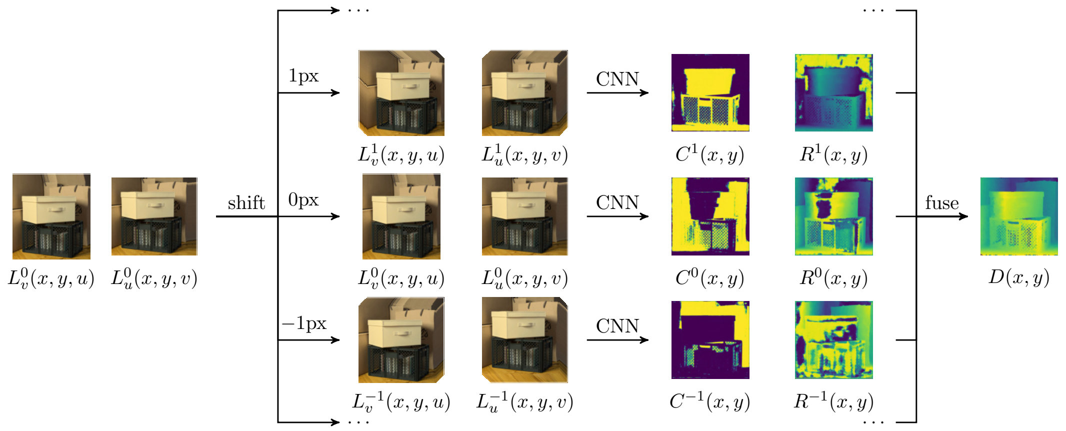


Figure 4.3: Method Overview. The input (left) consisting of two EPI stacks $L_v^0(x,y,u)$ and $L_u^0(x,y,v)$ is shifted several times, producing stacks with different disparity ranges. Our CNN (center) processes the shifted stacks, inferring a classification output $C^s(x,y)$ and regression output $R^s(x,y)$. Each pixel of the final result $D(x,y)$ (right) is assigned to a discrete disparity (classification) and refined by a sub-pixel disparity offset (regression).

4.2 Method

The core idea of our method is to use a plane sweep volume [73], [317] and successively apply the same neural network to each depth plane of that volume individually. The output of the network for each of these disparity ranges are two two-dimensional maps, one for the classification (correct plane or incorrect plane) and one for the disparity offset from the plane, in the range $[-0.5, 0.5]$. To generate the full disparity map of the scene, for each pixel, the shift with maximum classification activation is chosen to determine the correct plane. The corresponding per-pixel disparity offset is added to achieve sub-pixel accuracy. Because we are using a cross light field setup, the plane sweep volume can be constructed using the EPI-Shift approach, which refocuses the image stack by applying a shear transformation.

4.2.1 Light Field Camera Setup

Goal of our method is an accurate per-pixel disparity reconstruction within the center view of a $9+8$ cross-shaped light field camera setup with a large baseline. We limit ourselves to this setup due to the versatile usability in real world scenarios, compared to a star, or full four-dimensional light field setup. Many recent submissions to the HCI 4D Light Field Benchmark [10] demonstrate that research is shifting towards using more views from the available 81 input views, e.g. by synthesizing a focal stack from all views. However, we argue that this is a symptom of *benchmark optimization*, because adding more views gives diminishing returns and using less views is more practical in real-world applications. Note that the best approach by Schilling *et al.* [71] is not learning-based and requires only 17 views, compared to the inferior EPINET-Star [20] setup which requires nearly double the amount of views.

Four-dimensional light fields are recorded by a light field camera, arranged on a regular two-dimensional grid, indexed by $(u,v)^\top$. The baseline represents the distance between two adjacent cameras. Each camera captures a two-dimensional image with pixels, indexed by $(x, y)^\top$. A change of viewpoints on the u-axis for example, causes movement of a projected object point along the x-axis. The straight lines in image space, which represent this depth dependent movement, are called epipolar lines. For a cross-shaped light field, the two-dimensional slices of the light field along the xu- and yv-planes represent the EPIs. The x- or y-distance between the same object point in two adjacent views is the disparity d, measured in pixels and being inversely proportional to the depth. During camera calibration and rectification, the images often get pre-shifted. Therefore, also negative disparity values occur in some light field datasets. Refer to section 2.4 for more information.

4.2.2 EPI-Shift

Our EPI-Shift approach, which generates the plane sweep volume, boils down to a shear transformation on the xu-plane. Given an EPI-stack,

$$L_v^0(x, y, u) \tag{4.1}$$

defines the color value at a given image position $(x, y)^\top$, recorded by a camera u. The central camera is defined to be at $u = 0$. Positive u-indices are assigned to cameras that are located to the right of the central camera. As visualized in fig. 4.2, we perform the EPI-Shift by a certain disparity s with

$$L_v^s\left(x, y, u\right) = L_v^0(x - us, y, u). \tag{4.2}$$

Note, that in our notation L^s, the superscript s is not an exponent, but the amount of pixels that the light field is shifted with. Also note, that this operation refers to horizontal EPIs only. However, vertical EPIs $L_u^s\left(x, y, v\right)$ behave analogously after a rotation by 90° around the v-axis.

Because s is defined to be an integer number and we use a cross-shaped setup, no interpolation is required. We perform nearest-neighbor padding by clipping $x - us$ to the valid pixel range. However, in order to not waste capacity of the neural network for learning to deal with image borders, we do not apply any loss in areas affected by the padding.

To enable wide-baseline light field depth estimation, we perform three basic steps, illustrated in fig. 4.3. First, we generate a plane sweep volume by applying the EPI-Shift to the input light field, once per integer disparity within the disparity range of the scene. Each of those shifts can be thought of as a discrete disparity label. A pixel is assigned to a certain label if the disparity lies within $[-0.5, 0.5]$, when shifted by the labels disparity. Second, we infer a classification and a regression map for each of the shifts, using the CNN architecture described in section 4.2.3. Third, we compute the final result $D\left(x, y\right)$ for each pixel $(x, y)^\top$ in the center view by assigning an integer disparity label

$$D_{\text{int}}\left(x, y\right) = argmax_s\left(C^s\left(x, y\right)\right) \tag{4.3}$$

according to the shift s that produced the highest classification output $C^s\left(x, y\right)$. Using the regression map $R^s\left(x, y\right)$ of the respective shift, we add fine-grained disparity information to achieve the sub-pixel accurate result

$$D\left(x, y\right) = D_{\text{int}}\left(x, y\right) + R^{D_{\text{int}}(x,y)}(x, y). \tag{4.4}$$

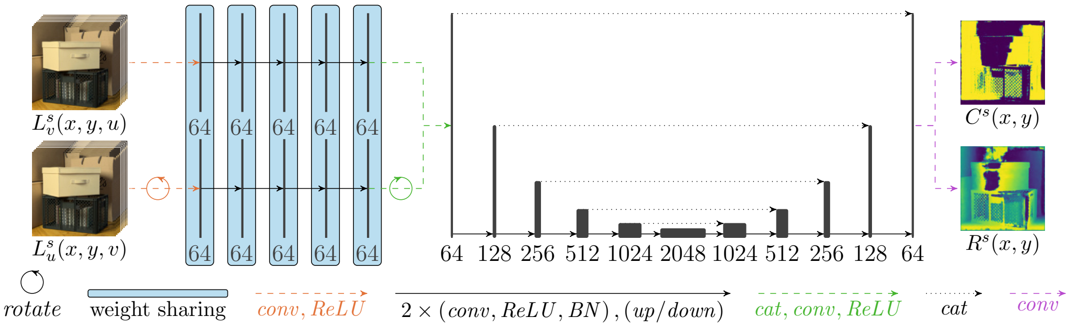


Figure 4.4: The neural network architecture of our model, consisting of two parts. First, a siamese feature extractor [20] (left), with four convolutional blocks for the discovery of local disparity information. Second, a U-Net architecture [90], [301] (right) to integrate global information. The input consists of two view stacks. Our network outputs a classification of discrete depth labels and a sub-pixel accurate disparity regression. The network solely uses convolutional blocks, consisting of two consecutive 3×3 convolutions with stride and padding of one. Numbers refer to the number of channels.

4.2.3 Network Architecture

Our architecture consists of two parts visualized in fig. 4.4. First, a siamese feature extraction network similar to Shin *et al.* [20]. The purpose of this subnetwork is the extraction of local disparity information. Second, a U-Net architecture (compare [301] and [90]) with two outputs: A classification output, assigning discrete per-pixel disparity labels and a continuous regression output, representing the sub-pixel accurate disparity relative to the label.

Siamese Feature Extraction Network: The cross-shaped light field provides a horizontal and a vertical stack of input views. Instead of concatenating the two, we chose a siamese twin architecture, consisting of one subnetwork for each stack. As both stacks contain similarly aligned EPIs after rotation of one stack by 90°, we share weights between the two subnetworks. This reduces the number of network parameters and therefore improves generalization.

The feature extraction network contains four fully-convolutional blocks, each consisting of two 3×3 convolutions followed by a *ReLU* activation function and a Batch Normalization (BN) layer each. We chose a number of 64 channels and preserve the input dimensions with a padding and stride of one.

To facilitate the classification for the downstream U-Net, we provide it with additional data. We apply the feature extraction network to both adjacent EPI-Shifts $L^{s\pm1}$. The extracted features are concatenated with those, extracted from L^s as well as the color information of the center view $L_{u=0,v=0}$. Hence, the number of input channels for the U-Net is, for each shift: The number of channels from the feature extraction subnetwork times two (horizontal + vertical) times three for the three shifts $\{-1, 0, +1\}$ plus the center view, i.e. a total of $64 \cdot 2 \cdot 3 + 3 = 387$ channels. Our experiments showed that these additional inputs improved the distinction between foreground and background objects in ambiguous regions. This is probably due to the depth hints from adjacent shifts that provide additional information about occlusions and therefore simplify classification. The addition of the center view allows the joint model to focus on feature extraction in the first part of the network, but still use the center view to guide smoothing in the U-Net part.

U-Net: A U-Net [301] architecture expands the effective receptive field of the joint model without loss of generalization capability. It therefore significantly improves the smoothness of non-textured areas. We chose a depth of five down- and upsampling layers leading to a receptive radius of 124 pixels for the U-Net part and 135 pixels for the whole network. The concatenated output of the upstream feature extractor network is reduced from 387 to 64 channels by an additional 3×3 convolution. For the processing inside the U-Net, we chose the same convolutional blocks as for the feature network, followed by an additional 3×3 up- or down-sampling convolution. The downsampling layers bisect the image dimensions while doubling the number of channels. Prior to upsampling, the output of the corresponding downsampling is concatenated. Therefore, the upsampling process doubles the image dimensions but divides the number of channels by four, please see Ronneberger *et al.* [301] for more details. A final 3×3 convolutional layer transforms the 64 output channels of the U-Net to two channels for the classification and regression output. Because the regressed disparity can be negative, no final ReLU activation function is applied.

4.2.4 Loss Function

Due to the drastically different outputs of our network, the choice of a well performing loss function is not trivial. For the classification output, a slight overlap between adjacent shifts might not have an effect on the final result at all. A large output at distant disparity regions, however, may cause a misclassification and therefore can destroy the end result. Our classification loss therefore specifically penalizes those cases. We define the loss with $C^s(x, y)$ being the classification output for a given shift s at pixel $(x, y)^\top$ as

$$\mathcal{L}_{class} = \sum_{s,x,y} \left(C^s(x, y) - C^s_{\text{true}}(x, y)\right)^2 \cdot \mathcal{W}_{disp}(x, y) \tag{4.5}$$

with a disparity weighting of

$$\mathcal{W}_{disp}(x, y) = \left(D(x, y) - D_{\text{true}}(x, y)\right)^2 \tag{4.6}$$

computed using the final disparity output $D(x, y)$ and the ground truth disparity $D_{\text{true}}(x, y)$ that penalizes misclassifications during training. The classification ground truth $C^s_{\text{true}}(x, y)$ should be high for all pixels within a disparity of ± 0.5 pixels. We therefore tried two different definitions: First, the one-hot or rectangle function

$$C^s_{\text{true}}(x, y) = \begin{cases} 1 & \text{if } |D_{\text{true}}(x, y) - s| \leq 0.5 + \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

producing hard boundaries between two labels. Second, the triangle function

$$C^s_{\text{true}}(x, y) = \max\left(0.5 + \epsilon - |D_{\text{true}}(x, y) - s|, 0\right) \tag{4.8}$$

which is more closely related to the regression output. It therefore should accelerate training and engage the network to share weight capacities between the two. On the other hand, it outputs lower values at boundaries, which are more vulnerable to misclassifications. In both cases, we choose a small ϵ that produces a slight overlap at the border regions between two disparity labels in order to prevent wrong classifications. We will see in the experiments that the rectangle function, defined in eq. (4.7), performs slightly better.

The regression output requires smooth surfaces but sharp edges. We see in our experiments that the $\mathcal{L}_1$ loss function fulfills those requirements for the regression loss best. We therefore define it as

$$\mathcal{L}_{reg} = \sum_{s,x,y} \left| R^s\left(x,y\right) - D_{\text{true}}\left(x,y\right) + s \right| C^s_{\text{true}}\left(x,y\right) \tag{4.9}$$

with $R^s\left(x,y\right)$ being the regression output. We mask out all pixels outside the subpixel interval by weighting with the rectangle function $C^s_{\text{true}}\left(x,y\right)$ defined in eq. (4.7). In order to compensate for misclassifications at the boundaries of disparity labels, we choose a slightly higher ϵ than for the classification ground truth. Due to the fundamentally different trend of the losses during training, a weighting between the two is also important for computation of the overall loss

$$\mathcal{L} = \alpha \mathcal{L}_{reg} + \mathcal{L}_{class}. \tag{4.10}$$

The choice of a scaling factor α depends on various factors such as the disparity distribution of the training data.

4.2.5 Training

We trained our model on the 16 so-called *additional* scenes of the HCI 4D Light Field Dataset [10]. We implemented the model in PyTorch [318] and trained it for four days on three Nvidia TITAN X GPUs. As optimizer we chose Adam with a learning rate of 10^{-4} for the first 10000 iterations. For another 30000 iterations, we decreased the learning rate to 10^{-5} and fixed the learned BN parameters.

We apply a large variety of data augmentation, comparable to EPINET [20], including random color channel re-distribution, random brightness and contrast adjustments, random rotations by multiples of 90°, random scales between 0.5 and 1 and random crops to a patch size of 225 × 225. This patch size leverages the utilization of global information by the U-Net. Our training batches contain seven shifts of two stacks extracted from a single RGB light field ($7 \cdot 2 \cdot 3 = 42$ channels).

4.2.6 Refinement

When choosing the rectangle classification ground truth defined in eq. (4.7), an additional refinement step can be performed. In case

$$\max_s \left(C^s\left(x,y\right)\right) < 0.01, \tag{4.11}$$

meaning that no classification exceeds some small threshold, we assume that the chosen disparity label at $\left(x,y\right)^\top$ is the wrong choice. In order to smoothly fill this pixel, we apply a median filter to each classification output first.

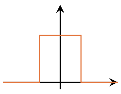
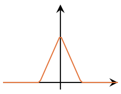
Class GT	$\epsilon = 0.0$ $\alpha = 0.25$	$\epsilon = 0.0$ $\alpha = 2.5$	$\epsilon = 0.25$ $\alpha = 0.25$	$\epsilon = 0.25$ $\alpha = 2.5$
	37.25	4.86	5.74	15.47
	37.43	52.22	24.40	5.27

Table 4.1: Mean Squared Error (MSE) score for the network, trained with different classification ground truth functions $C^s_{\text{true}}(x, y)$ and values for ϵ and α.

4.3 Experiments

In this section we present the results of our evaluations.

4.3.1 Ablation Studies

Because the choice of a classification ground truth function is highly important for our method, we evaluated different functions and parameters. Table 4.1 shows the MSE score of our network, trained on either the rectangle function, defined in eq. (4.7) or the triangle function, defined in eq. (4.8). As expected, the results show that the triangle function requires a higher ϵ to compensate for wrong classifications at boundaries of disparity labels. Due to the slightly better performance of the rectangle function, we chose it for our subsequent evaluations, setting $\epsilon = 0.17$ and $\alpha = 2.5$.

We also evaluated our CNN architecture, without EPI-Shift, similar to [20]. This model only reached an MSE score of 31.15 compared to 0.85 for our model with EPI-Shift. This clearly indicates that our U-Net architecture requires the shifted EPIs in order to properly generalize.

4.3.2 Results on the HCI 4D Light Field Benchmark

For the quantitative evaluation in fig. 4.6, we plot 13 error measures from the HCI 4D Light Field Benchmark. For details please refer to Honauer *et al.* [10]. Our method outperforms EPINET-Cross [20], in 11 out of 13 metrics with a close tie for the other two. Because our network is executed several times, once for each EPI-Shift, the runtime increases in comparison to EPINET [20]. However, our approach is still significantly faster than most classical optimization-based methods. Our work closes the performance gap to optimization-based methods while keeping all advantages of deep learning like fewer hyper parameters and learned instead of hand-crafted heuristics. We evaluated our method on four photo-realistic scenes of the publicly available HCI 4D Light Field Benchmark [10] that were not part of the training dataset. In fig. 4.5, we show a qualitative comparison with EPINET [20]. Note, that we use the cross setup for EPINET which uses the same 17 views subset of the full light field as is used by our approach. In addition to EPINET, the quantitative evaluation in fig. 4.6 also includes two state-of-the-art optimization-based methods

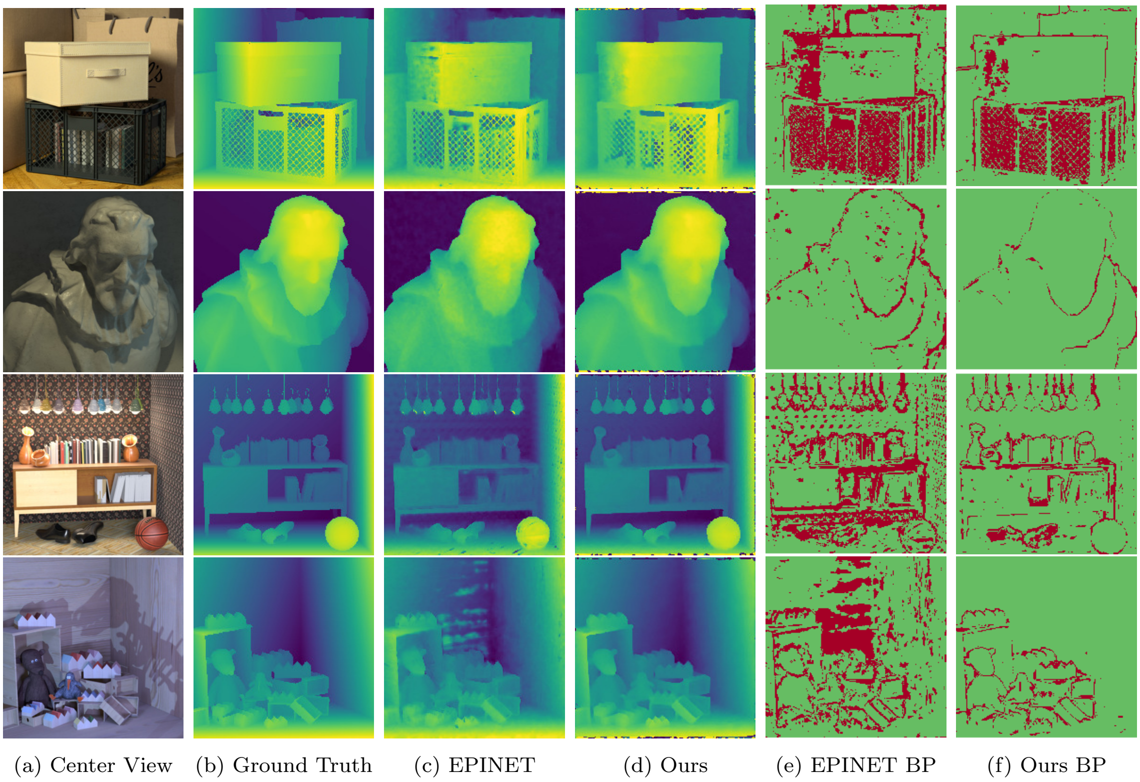


Figure 4.5: Results on the HCI 4D Light Field Benchmark [10] compared to the best learning-based competitor EPINET [20]. The BadPix score in (e) and (f) shows all pixels (red) exceeding an absolute distance of 0.07 to the ground truth.

(OBER [71] and SPO-MO [72]), ranking first and third in the official benchmark.

Our method provides considerably better quality at the disparity extremes (scene 3 and 4, background) due to the improved generalization enabled by EPI-Shift. Also note the improved performance on non-textured surfaces (scene 1, beige box) caused by the large receptive field of the U-Net which enables better smoothing and long-range reasoning. Unfortunately, the U-Net also seems to be more prone to noise at object boundaries which are not disparity label boundaries (compare scene 1, dark box), although similar artifacts can be observed with EPINET (compare scene 3, books). We refer to appendix A.1 for results on other benchmark scenes.

4.3.3 Results on Real Recordings

We also evaluated our method on images recorded with a real cross-shaped light field camera array with a disparity range of $[0, 12]$, consisting of 17 cameras. As the authors of [20] did not provide us with the pre-trained parameters for the cross-version of EPINET upon request, we trained EPINET-Cross based on their implementation. Figure 4.1 (left) shows one of the results. As expected, EPINET is only able to predict within the small training data disparity range of $[-3.5, 3.5]$, present in the background. In contrast, our EPI-Shift reconstructs the disparity in the whole range of $[0, 12]$.

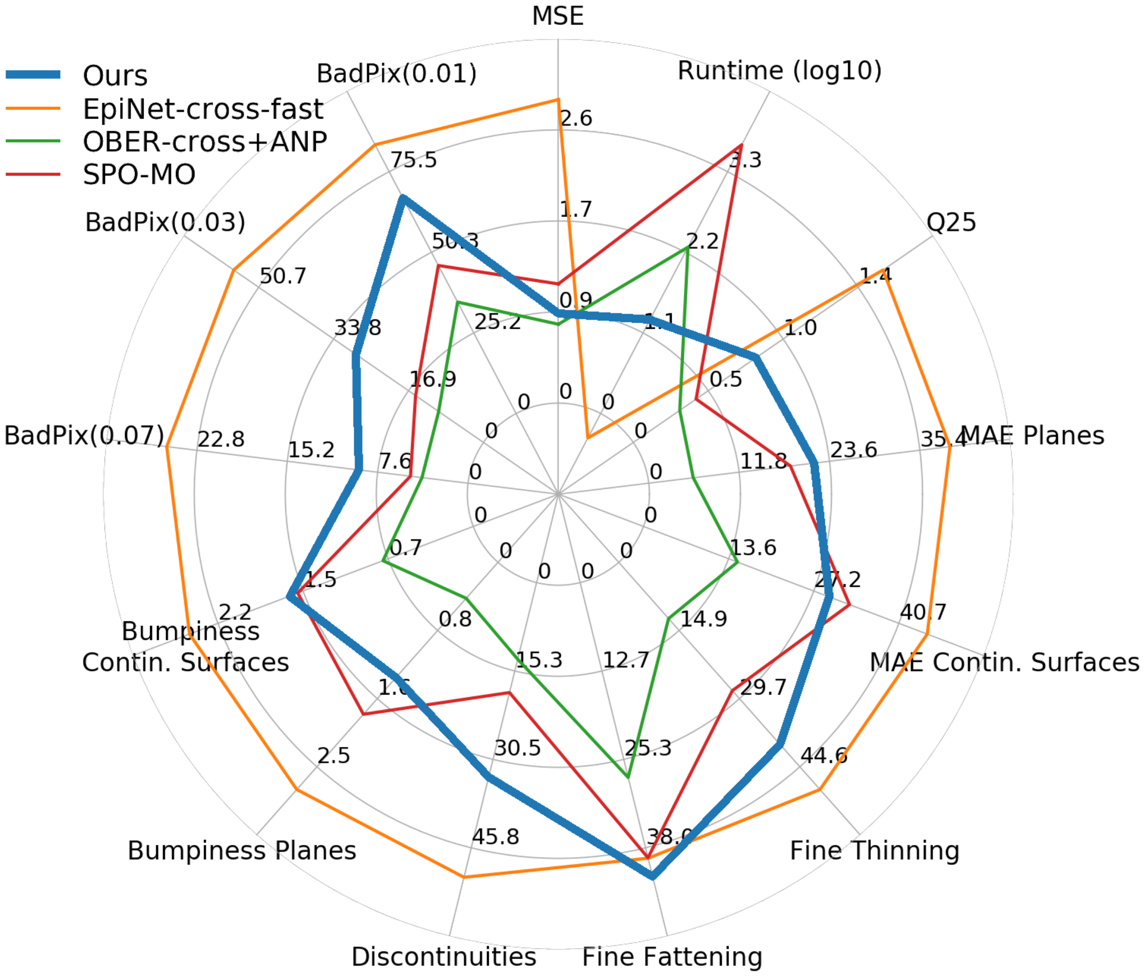


Figure 4.6: Qualitative results on synthetic data. We outperform EPINET-Cross [20] on 11 out of 13 metrics. Metrics and visualization provided by Honauer *et al.* [10].

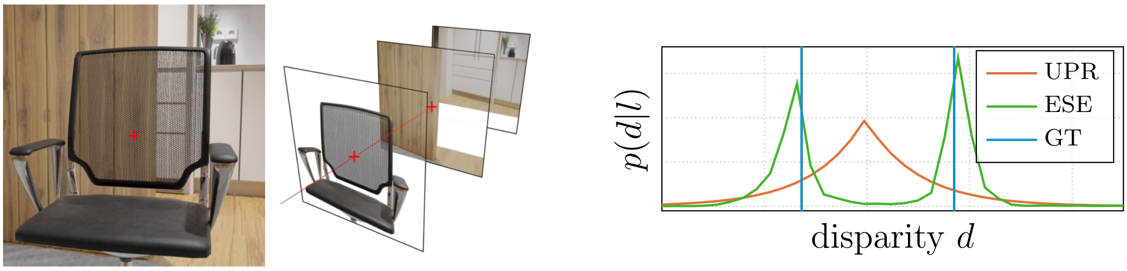
4.4 Conclusion

To summarize, we introduced a new learning-based approach for depth estimation from wide-baseline light field recordings. The key idea of our approach is to use EPI-Shifts, similar to plane sweep volumes for stereo depth estimation. This approach improves the generalization capability of CNN-based depth estimation and enables us to increase the receptive field using a U-Net which delivers better smoothing and reduces artifacts, thanks to long range reasoning. Combining these two advantages leads to state-of-the-art performance, as demonstrated on a publicly available light field benchmark. Furthermore, the EPI-Shift concept enables depth estimation for wide-baseline light fields, while the training data only exhibits small disparities. This greatly expands the range of possible applications. Previously, deep learning-based methods were mostly confined to light fields captured by plenoptic cameras. In contrast, EPI-Shift enables our approach to be used with wide-baseline, high-resolution camera arrays. These are often employed in industrial settings and for capturing training data for stereo and monocular depth estimation. To illustrate our method's versatility, we also present results from a wide-baseline real world scene.

5. Towards Multimodal Depth Estimation from Light Fields

Depth can have multiple modes if more than one three-dimensional point contributes to a light field pixel, which occurs at occlusions, with specular reflections, and on semi-transparent surfaces. This chapter introduces and analyzes various deep learning-based approaches to recover multiple depth modes.

5.1 Introduction



(a) Rendered scene (left) containing multiple disparity layers (right)

(b) Disparity posterior distributions, predicted by different methods

Figure 5.1: Comparison of disparity posterior distributions. Synthetic scene (a) containing overlapping objects at different depths. (b) shows disparity posterior distributions, estimated by different methods, for a single pixel (red crosses in (a)). This pixel captured two disparity modes (mesh material of chair (foreground) and wooden wall (background)). Note, that the Unimodal Posterior Regression (UPR) network, which outputs the mean and width of a Laplacian distribution, makes a wrong and uncertain prediction. The EPI-Shift Ensemble (ESE) detects both valid modes near the ground truth.

Light field recordings and their applications, like real time rendering for VR or highly accurate depth estimation, have improved vastly in recent years. However, while light field rendering methods handle transparent and reflective objects well, current depth estimation methods still perform poorly in those areas. State-of-the-art depth estimation methods mainly fail in three corner cases: at objects edges, semi-transparent and reflective surfaces. All three are caused by multiple objects at different depths contributing to the projected color of a single pixel on the camera sensor. Most existing models fundamentally ignore these cases and assume only one ‘true’ depth for each pixel.

Instead, we propose a series of deep learning-based methods to perform multimodal depth estimation, and depth estimation with uncertainty estimates. For this, we start with the basic idea of outputting Bayesian posteriors, whereas standard regression models just produce a single estimate. From this idea, and by using a simple and well-founded Maximum Likelihood (ML) training framework, we develop three different light field depth estimation methods, all three of which are able to infer uncertainty estimates, and use multiple ground truth values during training. Two of the proposed methods are also able to predict multiple distinct depth values per pixel at inference.

To train our methods, we propose to utilize a multi plane dataset that contains the exact depth, color and opacity of all depth planes that are visible in an image. This is, in contrast to other current datasets which only contain a single 'true' depth value. Our multi plane dataset consists of randomly generated synthetic scenes with a significant proportion of occlusion and transparent objects. This, for the first time, enables supervised training of multimodal depth estimation.

Our main contributions are as follows:

- An exploration of our three novel deep learning methods for light field based depth estimation being able to handle multiple depth modes: *(i)* Unimodal Posterior Regression (UPR); *(ii)* EPI-Shift Ensemble (ESE); *(iii)* Discrete Posterior Prediction (DPP).
- The release of the first multimodal light field dataset, containing the depth of all objects and their contribution to the color of each pixel in an image.
- A thorough evaluation of the predicted depth posterior distributions. We observe that the more restrictive UPR method works best when the unimodal depth assumption of traditional methods holds. However, in cases where it does not, the model is able to express a high uncertainty. In the context of multimodal areas the discrete DPP method is superior to ESE.

5.2 Method

Most methods for light field depth estimation assume opaque, smooth and Lambertian surfaces. Non-textured, specular or semi-transparent regions and depth edges are ambiguous and therefore challenging even for state-of-the-art methods. In this chapter we make progress for these cases by estimating the full depth posterior distribution. This is especially useful for pixels with more than one valid depth mode, caused by either semi-transparency or the point spread at depth edges. Unlike previous works that only predict a single depth, we are able to find those modes.

5.2.1 Posterior Estimation

The previous chapter used a precise notation with indices for image and view coordinates to accurately describe the EPI-Shift approach. This chapter requires a more probabilistic notation and therefore emits those indices where possible, for ease of understanding. From here on, we will simply refer to the input light field of a depth estimation network as l, and to disparity as d. In practice, the input to

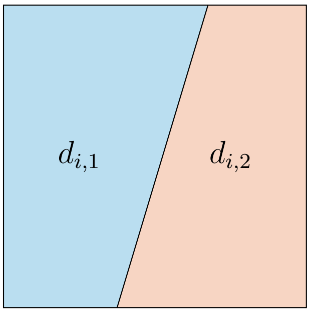


Figure 5.2: Visualization of opacity of two disparities $d_{i,1}$ and $d_{i,2}$ visible in the same pixel. Opacity, which is used in our definition, corresponds to the fraction of the area that an object takes up within a pixel before integration or rendering.

a depth estimation network is a concatenation of horizontal, vertical and diagonal EPIs. Standard regression models usually output a single guess for the disparity, $d = f_w(l)$, where f_w could be EPINET [20] with network weights denoted as w. Instead, our goal is to estimate the posterior distribution $p(d \mid l)$ of the disparity d given an input light field l. In the following, we present four different approaches that all model such a posterior distribution.

To supervise more complex posterior distributions that can represent more than one mode, we created our own multimodal depth dataset. Unlike common datasets that only contain a single ground truth disparity d_i for each pixel i, we include the disparities of multiple depth modes $d_{i,j}$ for transparent objects and depth edges. For each disparity, we also include the amount of color $\eta_{i,j}$ that it contributed to the pixel, i.e. the perceived opacity of that object in the pixel. From a Bayesian perspective, we interpret $\eta_{i,j}$ as the probability $p(d_{i,j})$ of this disparity. This choice is justified in both an intuitive and methodological sense: $\eta_{i,j}$ corresponds to the fraction of the pixel's area that is taken up by the object at disparity $d_{i,j}$. Lacking any prior knowledge, this is also equal to the probability that the depth at any subpixel position corresponds to that object. This equality between the opacity $\eta_{i,j}$ and probability $p(d_{i,j})$ is valid both at edges as well as fine structures such as grids or woven meshes. We also extend the definition to apply to semi-transparent materials such as printed glass as a simplifying assumption.

From a Bayesian perspective, the probability $p(d_{i,j})$ of each possible ground truth disparity value for a pixel quantifies the *degree of belief* in this value. For a synthetic dataset, in absence of a real ground truth measurement device whose characteristics we can analyze, any definition for $p(d_{i,j})$ is valid as long as it leads to stable training and a model that reproduces the different modes with their corresponding probabilities faithfully at test time.

However, there are still some choices which are more sensible or well founded than others. In terms of the opacity η_j, it should be evident to choose

$$\eta_j = 0 \implies p(d_{i,j}) = 0 \text{ and} \tag{5.1}$$

$$\eta_j = 1 \implies p(d_{i,j}) = 1, \tag{5.2}$$

meaning that if an object is not visible at all in a pixel, its disparity should not be considered, and vice versa, if an object is the only one visible in a pixel, its disparity should be the only valid answer. In between these two points, we argue for the simplest choice of $p(d_{i,j}) = \eta_j$. We note that if a setup requires a different definition of $p(d_{i,j})$, e.g. re-weight to increase the dominant mode, up-weight the foreground mode, etc. the posterior can easily be re-weighted at test time, without retraining the model. This is only possible with methods such as ours that produce a full posterior.

Despite various valid choices of defining $p(d_{i,j})$, we do argue that our definition makes practical sense: the opacity corresponds to the fraction of the area that an object takes up within a pixel before integration or rendering. It is therefore equal to the probability that the depth of that object would be observed when measuring at a random subpixel position. In other words, if we were to take many physical depth measurements within a pixel, the relative occurrence of each measured depth value $d_{i,j}$ (therefore arguably the probability $p(d_{i,j})$), would be the same as the opacity η_j. While this applies exactly to our synthetically rendered dataset, some additional effects such as point spread functions and non-uniform pixel integration functions would apply for real recorded light fields. These effects might make the derivation more complex, but do not change the general idea.

Unimodal Posterior Regression

The most common approach for learning distributions is ML learning, which most loss functions can be reformulated as. The ML objective aims to find the model parameters w which maximize the log likelihood of the training data $\{(l_i, d_i)\}_{i=0}^{N}$ under the estimated posterior distribution. In practice, we minimize the *negative* log likelihood instead:

$$\mathcal{L}_{\mathrm{ML}} = -\frac{1}{N} \sum_i \log p(d_i \mid l_i, w). \tag{5.3}$$

It can be shown that this objective minimizes the Kullback-Leibler Divergence (KLD) between $p(d \mid l, w)$ and the true posterior $p_{\mathrm{true}}(d \mid l)$.

Previous regression based approaches, like EPINET [20], simply use the $\mathcal{L}_1$ loss to make a single prediction:

$$\mathcal{L}_1 = \frac{1}{N} \sum_i \left| d_i - f_w(l_i) \right|. \tag{5.4}$$

We see that this is equal to the ML objective when the posterior is assumed to be a Laplace distribution $p(d \mid l, w) \propto \exp(-|d - \mu|/b)/2b$ with the network output $\mu = f_w(l)$ and a fixed value of $b = 1$. This motivates the following simple extension, which is an adaptation of the Dawid-Sebastiani score [319], later popularized by Kendall and Gal [313], except that the $\mathcal{L}_2$-loss corresponding to a Gaussian posterior was used instead: we allow the network to change the width b of the posterior. With this, it becomes

$$p(d \mid l, w) = \frac{1}{2b} \exp\left(-\frac{|d - \mu|}{b} \right), \text{ with } \quad [\mu, b] = f_w(l). \tag{5.5}$$

Putting this back into the ML objective, we get the following loss function for the predictive uncertainty:

$$\mathcal{L}_{\mathrm{UPR}} = \frac{1}{N} \sum_i \frac{|d_i - \mu_i|}{b_i} + \log b_i, \text{ with } \quad [\mu_i, b_i] = f_w(l_i). \tag{5.6}$$

This loss can be understood intuitively: If the network struggles to predict μ_i, the $\mathcal{L}_1$ loss term can be down-weighted by increasing the scale parameter b_i for this pixel. To avoid the trivial solution $b \to \infty$ for any input, high b are penalized by a regularization term $\log b$. In practice, we let the network predict $\log b$ instead of b to improve numerical stability.

This approach gives us a measure of aleatoric uncertainty [313] for each pixel which is already helpful for many downstream applications. However, the implicit assumption for the method to work well is that the true posterior is also Laplacian. Needless to say, this is certainly not true in multi-modal cases, which cannot be modeled by the Laplace distribution. With multiple ground-truth depth modes $d_{i,j}$ as opposed to only d_i, as in our dataset, the loss for a Laplace distribution becomes

$$\mathcal{L}_{\text{UPR}}^{\text{MM}} = \frac{1}{N} \sum_i \sum_j p(d_{i,j}) \frac{|d_{i,j} - f_w(l_i)|}{\log b_i} + \log b_i \tag{5.7}$$

and can be applied to the $\mathcal{L}_1$ loss respectively:

$$\mathcal{L}_1^{\text{MM}} = \frac{1}{N} \sum_i \sum_j p(d_{i,j}) |d_{i,j} - f_w(l_i)|. \tag{5.8}$$

However, in any case, those networks will focus on a single mode, or lie in between, and compensate for its wrong prediction by expressing a very high uncertainty like in fig. 5.1.

EPI-Shift Ensemble

Commonly, one way to circumvent this exact issue is to use an ensemble of networks. Instead of just estimating a single posterior, M networks predict M different posteriors, which are then averaged:

$$p(d|l) = \frac{1}{M} \sum_k p(d|l, w_k) \tag{5.9}$$

over all networks with learned weights w_k, $k = 1 \dots M$. It has been shown that ensembles deliver some of the best uncertainty estimates among existing methods [320]. Their main limitation is the high computational cost, especially for training. Various approaches try to avoid this and train only a single model. For instance, Monte Carlo dropout exhibits similar characteristics as a true ensemble [321]. Instead, we propose a new scheme which uniquely exploits the nature of light field data, which we term EPI-Shift Ensemble (ESE). Therefore, we extend the EPI-Shift operation, introduced in the previous chapter, and extend it to arbitrary sub-pixel steps Δd. The shift transformation allows us to apply a disparity offset s to any light field l. In contrast to the original method which only applies integer pixel shifts, we also need sub-pixel shifts to ensure the detection of modes that are closer than one pixel. To achieve this, we apply a linear interpolation. Thus the original formulation for a horizontal EPI

$$L_v^s(x, y, u) = L_v^0(x - us, y, u) \tag{5.10}$$

can be generalized to continuous s using linear interpolation

$$L_v^s(x, y, u) = \alpha L_v^0(\lfloor x - us \rfloor, y, u) + (1 - \alpha) L_v^0(\lceil x - us \rceil, y, u) \tag{5.11}$$

with an interpolation factor $\alpha = \mathrm{frac}(us)$. This can be adapted trivially to vertical EPIs. For diagonal EPIs, the horizontal and vertical shift is applied successively.

However, this operation alone does not prevent the problems seen for the single mode approach: the network will try to average out bi-modal solutions, or collapse into one mode. As a result, we see little to no diversity in the ESEs, and no multi-modal posteriors. To prevent the collapse, we mask the loss during training so that it only applies to pixels with $|d| < \Delta d/2$. In all other cases the output will have a large uncertainty:

$$\mathcal{L}_{\mathrm{ESE}} = \frac{1}{N} \sum_i \begin{cases} \frac{|\mu_i - d_i|}{b_i} + \log b_i & \text{if } |d_i| < \frac{\Delta d}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

We extend this to our multimodal dataset similarly to our unimodal networks:

$$\mathcal{L}_{\mathrm{ESE}}^{\mathrm{MM}} = \frac{1}{N} \sum_i \sum_j p(d_{i,j}) \begin{cases} \frac{|\mu_i - d_{i,j}|}{b_i} + \log b_i & \text{if } |d_{i,j}| < \frac{\Delta d}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{5.13}$$

After training with this loss, the network will only be confident (narrow posterior) if it estimates that the input disparity is $d \in [-\Delta d/2, \Delta d/2]$, and the predicted posterior will always be centered in this range. When using this network in the ESE, we see that each term k will only contribute a narrow posterior if a plausible disparity lies between $(k - 1/2)\Delta d$ and $(k + 1/2)\Delta d$, thus ensuring diverse outputs and the possibility of multi-modal predictions. Three details should be noted: First, the model is not trained as an ensemble, a single model is trained just as before with the modified $\mathcal{L}_{\mathrm{ESE}}$-loss. The ensembling operation is only performed at inference time. Second, the masked loss does not reduce the effective size of our training set, as we also apply random EPI-Shifts as a part of the data augmentation process. This way, all pixels will randomly fulfill $|d_i| < \Delta d/2$ at some point. Lastly, the inference time is M times longer, as M forward passes have to be performed to compute the ESE.

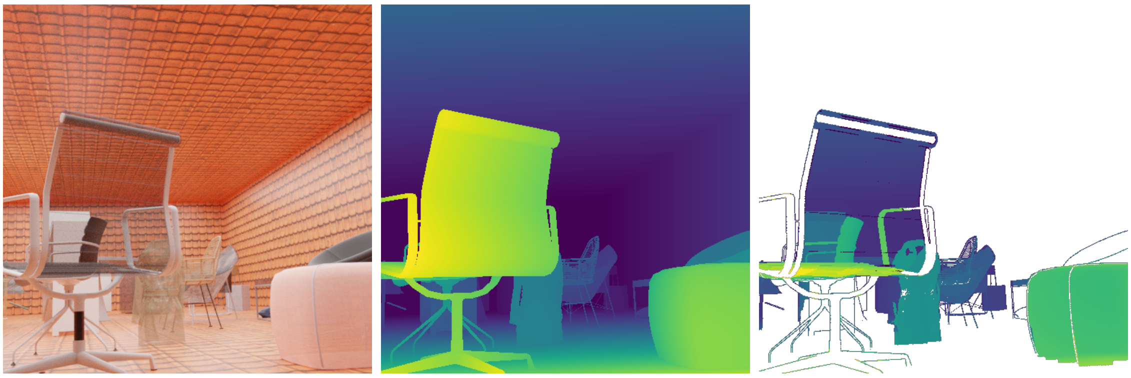
Discrete Posterior Prediction

The approach of discretizing regression tasks has been successful for stereo depth estimation in the past and promises to model more expressive posteriors. Specifically, by discretizing the range of disparities, a softmax output can be used to represent the posterior. The posterior is then a step function consisting of these discrete probabilities. If d_j are the discretization steps, we can write

$$p(d_j|l) \propto \mathit{softmax}\left(f_w(l_i)\right)_j := \frac{\exp\left(f_w(l_i)_j\right)}{\sum_k \exp\left(f_w(l_i)_k\right)}. \tag{5.14}$$

If multiple modes are used for the training, we also discretize the distribution $p(d_j)$ over these modes. Maximum likelihood training is then simply equivalent to the categorical Cross Entropy (CE) loss, where the correct 'class' is the bin j that the training example d_i lands in:

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N} \sum_i -\log\left(\mathit{softmax}\left(f_w(l_i)\right)_j\right). \tag{5.15}$$



(a) Exemplary scene (b) First disparity (c) Second disparity

Figure 5.3: An exemplary scene (a) from our randomly generated dataset. (b) and (c) show different disparity modes. Each pixel has at least one disparity mode. Behind semi-transparent objects and at depth edges, a second disparity mode (c) exists.

For the multimodal dataset, we can compute the CE accordingly:

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{MM}} = \frac{1}{N} \sum_{i} \sum_{j} -p(d_j) \log \left(\mathit{softmax} \left(f_w(l_i) \right)_j \right). \tag{5.16}$$

Note, that for a unimodal dataset, we simply have $p(d_j) = 1$ and $p(d_{l \neq j}) = 0$, simplifying to the CE in eq. (5.15).

Although the discrete posterior prediction gives us an uncertainty estimation at much lower computational cost than the ESE and can represent more flexible posteriors compared to simple Laplacians, softmax probabilities are generally known to be overconfident [322] and also make wrong but confident predictions in ambiguous or unseen cases. Different techniques for post-calibration of uncertainties exist, outlined in [322]. However, while they may prevent overconfidence going from the training to a test set, they do not make the uncertainties more reliable in general [320], e.g. for ambiguous inputs.

5.2.2 Dataset Generation

To train and validate all methods above, ground truth multimodal depth data is required. Because all previous light field datasets only contain a single depth per pixel, we generated a novel multimodal depth light field dataset containing 110 randomly generated indoor scenes. To improve the training performance, the dataset generator follows four goals: *(i)* relatively photorealistic appearance *(ii)* high diversity to improve generalization of trained models *(iii)* many occlusions and depth edges to improve the performance at object edges and *(iv)* a large proportion of pixels with multiple valid depths. To maximize the occlusions, we generate relatively deep indoor room scenes with a high number of objects. From a set of ca. 750 assets, mainly furniture and accessories, we randomly choose 48 objects per scene and place them in a non-colliding way on the floor. In addition, random materials with a random opacity are chosen to increase the number of semi-transparent surfaces. To maximize the diversity, we also randomly choose one of 750 tileable textures for the walls, ceiling and floor. We then render the created scene by separating it into 128

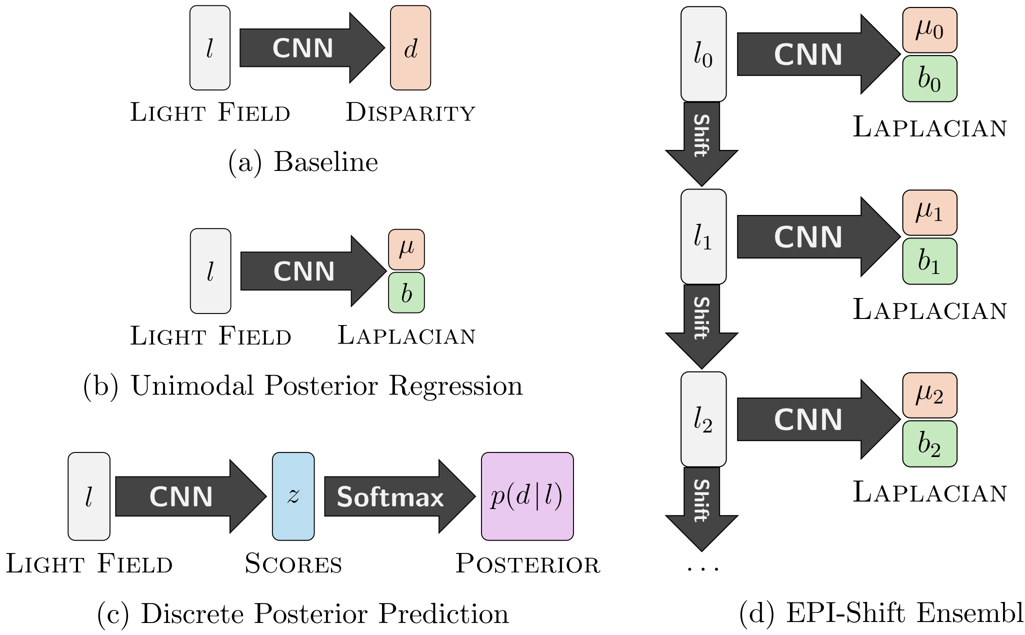


Figure 5.4: Network architecture overview. Our baseline is a simple feed-forward network trained to only predict disparity (a). We compare three posterior-regression methods: Laplacian prediction using learned loss attenuation (b), an ensemble on shifted inputs (d) and a discrete softmax classification network (c).

slices of equal depth, because we observed that this leads to different objects falling into different slices almost always. We then render the color, alpha transparency and depth of each pixel for each slice. Alpha compositing follows the *over operator*

$$c_0 = \frac{c_1\alpha_1 + c_2\alpha_2(1-\alpha_1)}{\alpha_0} \tag{5.17}$$

with c_0 being the resulting color from color c_1 rendered *over* color c_2. The new alpha opacity of color c_0 is

$$\alpha_0 = \alpha_1 + \alpha_2(1-\alpha_1). \tag{5.18}$$

The contribution $p(d_j) = \eta_j$ of the color c_j at disparity d_j is therefore calculated as

$$p(d_j) = \eta_j = \alpha_j\left(1 - \alpha_{j-1}\left(1 - \alpha_{j-2}\left(1 - \ldots \alpha_0\right)\right)\right). \tag{5.19}$$

Lastly, we save all depths for each pixel that are not fully occluded.

5.3 Experiments

In this section, we describe our training procedure and analyze the predicted posterior distributions. We therefore distinguish two possible applications: unimodal prediction with uncertainty and multimodal prediction.

5.3.1 Architectures and Training Details

To ensure a fair comparison, we chose the state-of-the-art method EPINET [20] with minor modifications as our backbone network architecture for all models. We describe the exact network architectures in more detail in appendix B.1. The network consists of a total number of eight blocks, containing two convolutional layers, followed by a ReLU each and one BN layer. For UPR, we added an additional output layer to also predict the negative log width $-\log b$. The ESE is a modification of UPR with the only difference being the loss functions. Our extended EPI-Shift transformation is described in more detail in chapter 4. For the discrete DPP method, we chose a number of 108 'classes'. This is motivated by the common BadPix007 metric that considers a pixel as correct if it is closer than 0.07px to the ground truth. A number of 108 classes in a disparity range of $[-3.5, 3.5]$ leads to a bin size of ≈ 0.065 which is slightly below this threshold.

We trained all networks using the loss functions described in section 5.2.1. The unimodal loss functions, denoted as $\mathcal{L}_x$ are always applied to the closest disparity. All multimodal loss functions, denoted as $\mathcal{L}_x^{\text{MM}}$ are applied to all disparity modes.

In addition, we reimplemented EPINET [20] in our own framework as a baseline for a fair comparison. The learning rate was set to 10^{-3} for all models, using a batch size of 512 and the Adam optimizer. We trained on randomly cropped patches ($96px \times 96px$) from the set of 100 training scenes in our dataset. To further improve the diversity of our dataset, we make use of a number of data augmentation operations: We apply a random sub-pixel EPI-Shift in the range $[-2, 2]$. In addition, we randomly rotate the light field by multiples of 90°, randomly rotate the colors in RGB-space and randomly change brightness and contrast.

5.3.2 Posterior Evaluation

Depth estimation methods are usually evaluated by measuring the pixel-wise error to a ground truth disparity map. However, to correctly measure the quality of estimates in areas with multiple valid depths, a different set of metrics is required. We consider two application scenarios:

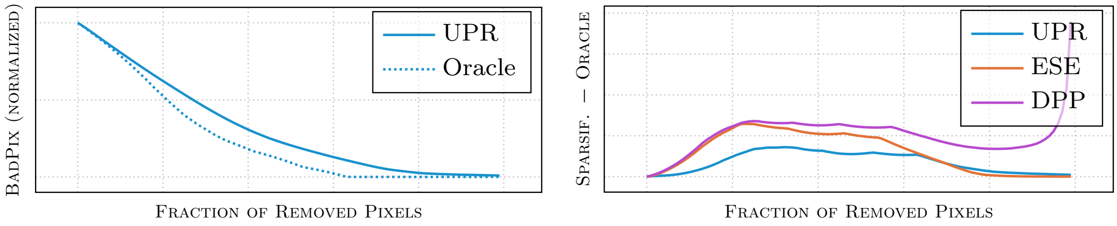
First, the estimation of just a single disparity, but with an additional confidence measure to ensure that the estimate can be trusted. This may be required, e.g. in industrial and robotics applications where decisions are based on the estimated depth.

Second, the estimation of multiple depths in areas with transparent objects or at object edges. This is typically required by computer graphics applications that aim to render the recorded scene from a different angle.

In the following, we introduce metrics for both cases.

Unimodal Prediction with Uncertainty

Previous methods and datasets always consider the disparity of the closest object as 'true', even when this object is transparent. However, estimation methods oftentimes output the disparity of the background object in those cases, which may lead to severe issues in downstream applications. In addition, more ambiguities usually occur in non-textured areas which cannot be estimated correctly. Most optimization-based methods fill in those ambiguous regions by interpolation between adjacent pixels with confident predictions. Due to the limited receptive field of neural network based methods, this is only possible to some extent. In any case, a confidence measure is extremely useful for downstream applications in order to decide if an estimate can be trusted. To achieve this, the overall variance σ^2 of



(a) Sparsification curve

(b) Sparsification Error (SE)

Figure 5.5: Unimodal uncertainty quantification. Sparsification results of analyzed methods with respect to the disparity BadPix007. By removing a certain fraction with the highest predicted uncertainty (a), the error decreases. The *Oracle* is a lower bound, created by removal of truly worst pixels. We also compute the difference between the predicted sparsification and its *Oracle*, denoted as SE (b), for a comparison of the three methods, trained on all depth modes.

Method	**Unimodal Metrics**		**KLD**			**AuSE** ↓	**Time** ↓
	MSE ↓	BadPix ↓	Unimodal ↓	Multimodal ↓	Overall ↓		(in sec)
BASE (uni)	**0.374**	0.229	4.720	7.876	5.421	-	**2.188**
BASE (multi)	0.563	0.307	5.259	8.514	6.025	-	2.211
UPR (uni)	0.439	0.235	1.719	3.381	1.879	**0.071**	2.260
UPR (multi)	0.676	0.285	1.987	3.156	2.114	0.072	2.287
ESE (uni)	1.269	0.223	4.164	3.628	4.160	0.099	17.492
ESE (multi)	1.850	0.229	4.283	3.719	4.277	0.121	16.902
DPP (uni)	0.765	**0.209**	**1.631**	3.057	**1.734**	0.272	4.348
DPP (multi)	0.686	0.231	1.824	**2.987**	1.914	0.197	4.382

Table 5.1: Evaluation, from left to right: MSE and the common BadPix007 score (percentage of pixels with $|d - d_{\text{true}}| > 0.07$), KLD divergence on unimodal, multimodal and all pixels, AuSE, runtime of one forward pass. Our methods were trained with both losses $\mathcal{L}_x$ (unimodal) and $\mathcal{L}_x^{\text{MM}}$ (multimodal) respectively. Lower is better.

the predicted posterior distribution can be used as an uncertainty measure. We aim for a consistently high uncertainty in regions with ambiguous predictions. To evaluate the quality of the estimated posteriors, we remove the $x\%$ of pixels with the highest posterior variance (uncertainty). With these ambiguous and uncertain cases filtered out, the BadPix of the remaining pixels is lower, which can be plotted as a sparsification curve (see fig. 5.5a). The optimal curve can be computed by removing those $x\%$ of pixels with the largest ground truth error. We call this the *Oracle* curve, which represents the lower bound of what is achievable. This method is commonly used to evaluate uncertainties for regression tasks [314]. In order to compare all methods, we compute the Sparsification Error (SE) by subtracting the *Oracle* curve from the sparsification curve. The Area under the Sparsification Error (AuSE) quantifies the uncertainty quality of each method with a single number.

Multimodal Prediction

For many applications, including re-rendering of a recorded scene from different angles, estimating multiple depths for pixels at object edges and transparent surfaces is desirable. In addition to the disparities of all modes, the contribution of each mode to the color of the pixel is also important. To evaluate both, we measure the KLD

$$\mathcal{D}_{\text{KL}} = \int p(d_i) \log \left(\frac{p(d_i)}{p(d_i|l_i)} \right) \tag{5.20}$$

between the predicted disparity posterior $p(d_i|l_i)$ and the true disparity distribution $p(d_i)$ at a pixel i. Intuitively, the KLD will be minimal if the posterior assigns a high probability density to each true disparity mode and a low density to disparities that are not present at a pixel i. Optimally, the density at each disparity mode corresponds to the contribution η_i of this mode to the resulting pixel color. As the KLD is only well defined between two continuous or two discrete distributions and we compare continuous as well as discrete methods, we chose to discretize all distributions. We therefore assign each ground truth disparity to one out of $K = 108$ bins with a width of $h \approx 0.065$px each.

$$p(d_k) = \sum_j p(d_j) \; \forall j \text{ with } |d_j - d_k| < \frac{h}{2} \tag{5.21}$$

This is, again, motivated by the well-established BadPix007 metric [10] which considers a pixel as correct if the $\mathcal{L}_1$ distance to the ground truth is below 0.07px. For the baseline method that only outputs one disparity, we simply set the probability of the bin that contains this disparity to one. All continuous posterior distributions are discretized by integrating over the interval of each bin:

$$p(d_k|l) = \int\limits_{(k-0.5)h}^{(k+0.5)h} p(d|l) \mathrm{d}d \tag{5.22}$$

We now average the discrete Kullback-Leibler Divergence (KLD)

$$\mathcal{D}_{\text{KL}} = \frac{1}{NK} \sum_i \sum_k p(d_{ik}) \log \left(\frac{p(d_{ik})}{p(d_{ik}|l_i)} \right) \tag{5.23}$$

over all pixels in all validation scenes. In addition, we also compute the KLD over all unimodal and all multimodal pixels separately. A pixel is considered multimodal if it has at least two modes j with $p(d_j) > 0.3$.

Results

Table 5.1 compares the unimodal, multimodal and sparsification performance of all methods. In the following, we will interpret our results with respect to the aforementioned applications: unimodal disparity estimation with uncertainty and multimodal disparity estimation.

When considering pure unimodal performance, our baseline method and DPP perform best. The higher MSE for DPP is caused by small discretization errors due to the discrete number of bins. Those small errors are well below a threshold of 0.07px and therefore ignored by the BadPix metric which shows that DPP indeed predicts 2% more pixels correctly compared to the baseline. UPR performs only slightly worse than both methods overall. However, due to the uncertainty being directly supervised by $\mathcal{L}_{\text{UPR}}$ this method is superior in terms of sparsification. This

Method	**Unimodal Metrics**		**KLD**			**AuSE ↓**	**Time ↓**
	MSE ↓	BadPix ↓	Unimodal ↓	Multimodal ↓	Overall ↓		(in sec)
BASE (multi)	**0.435**	0.274	4.807	8.081	6.078	-	**0.557**
UPR (multi)	0.480	0.285	2.028	3.551	2.448	**0.115**	0.578
ESE (multi)	1.204	0.245	4.330	3.769	4.226	0.182	4.502
DPP (multi)	0.608	**0.239**	**1.786**	**3.193**	**2.136**	0.288	1.068
SLFC [123]	3.449	0.660	3.694	3.908	3.715	0.324	1054.231

Table 5.2: Comparison to Sparse Light Field Coding (SLFC) [123], from left to right: Mean Squared Error and the common BadPix007 score (percentage of pixels with $|d - d_{\text{true}}| > 0.07$), KLD divergence on unimodal, multimodal and all pixels, AuSE, runtime of one forward pass. Our methods were trained using the multimodal loss $\mathcal{L}_x^{\text{MM}}$. Lower is better.

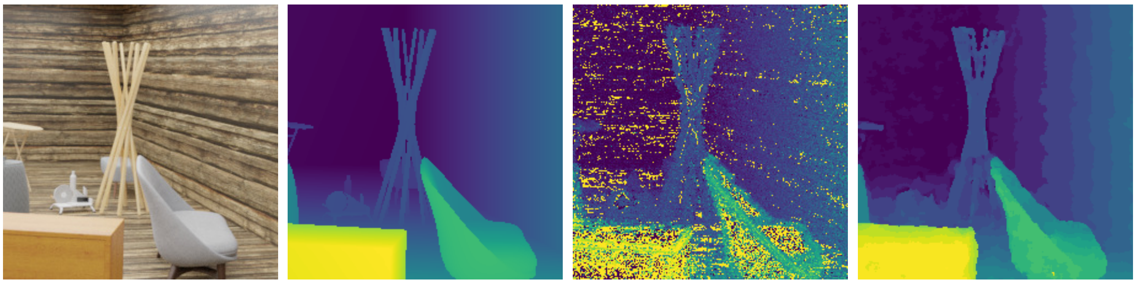
means that its uncertainty metric reflects most accurately whether a prediction is correct. We conclude that, if the application requires only a single disparity and confidence is important, UPR should be considered.

With respect to the accuracy of the predicted posterior distribution, DPP performs best in unimodal and also multimodal areas. However, as most softmax prediction methods, it is overconfident, as reflected by the SE in fig. 5.5b. Despite the popularity of ensemble-based models for uncertainty estimation, ESE cannot compete with the other two methods. We observe that in generally ambiguous (non-textured) areas, the uncertainties, estimated by each 'ensemble member' of ESE are very similar. Therefore, a seemingly random disparity from the whole disparity range is chosen, while the predictions of all other networks are usually smooth, even in those uncertain areas. This is also reflected by the relatively high MSE but low BadPix error: the amount of 'correctly' predicted pixels is on par with other methods, but the deviation of 'wrongly' predicted pixels is generally higher. In addition, as all members contribute slightly to the mixture of Laplacians, the density of the posterior is higher along the whole disparity interval, which leads to a worse multimodal KLD compared to UPR and DPP. We therefore generally recommend DPP for multimodal predictions in small, narrow-baseline light fields. However, due to its shift-operation, ESE can, unlike other methods, operate on arbitrary large disparity ranges and is therefore still advisable for high-resolution or wide-baseline light field cameras. In addition, it performs also relatively well in terms of sparsification. Comparing the methods trained on only one mode with the same methods trained on multiple modes shows that multimodal training leads to a slightly better multimodal performance for UPR and DPP, but always comes at a cost in unimodal areas. Our baseline method cannot efficiently represent multimodal posteriors as it only predicts a single disparity.

We conclude that the exact model and training method should be carefully chosen based on the intended application.

5.3.3 Comparison to Sparse Light Field Coding

We also compared our methods to "What Sparse Light Field Coding Reveals about Scene Structure" by Johannsen *et al.* [123]. This is, to the best of our knowledge, the only previous method which is able to estimate multiple depth modes. The method uses a dictionary of small EPI patches. Each atom in this dictionary corresponds to a unique disparity. On small EPI windows around each pixel, the Lasso optimizer



(a) Center view (b) Ground Truth (c) SLFC [123] (d) DPP

Figure 5.6: Qualitative results of SLFC [123], compared to DPP on one of our multimodal validation scenes: We chose the disparity which corresponds to the strongest coefficient for each pixel. Compared to our deep learning-based methods, SLFC [123] tends to wrong classifications in non-textured areas which causes noise. This also has a negative impact on SLFCs posterior prediction performance.

is used to infer the coefficients for each atom. A large coefficient for an atom means that the disparity which corresponds to this atom was observed at this pixel. The vector of coefficients can therefore also be interpreted as a discrete disparity posterior distribution, similarly to DPP. The authors were able to provide us with only a part of the code which we used to create the dictionaries for our multimodal validation dataset. For a fair comparison, we again chose a number of 108 disparity steps. We used the Lasso optimizer from the Python framework *scikit-learn* and set $\alpha = 0.01$ as recommended by the authors. After optimization of the posterior distribution for each pixel, we compared the method to our four deep learning-based models. Please note, that due to the enormous runtime of SLFC (even with our parallel implementation on 128 Central Processing Unit (CPU) cores), we run it on a cropped (0.5×0.5) version of our validation dataset. For a fair comparison, we ran the methods trained on the multimodal posterior distribution with loss functions $\mathcal{L}_x^{MM}$ on the same cropped scenes.

Table 5.2 shows the results of our comparison. We notice that SLFC produces more wrong classifications in non-textured and therefore uncertain areas which leads to more overall noise. We argue that this is due to the local per-pixel optimization. In contrast, our neural networks benefit from a larger receptive field and are therefore capable to deliver smooth results, even within relatively large non-textured areas (compare fig. 5.6). This effect causes an overall worse performance of SLFC. To compute the unimodal metrics, we chose the discrete disparity with the highest posterior probability for each pixel. Both, the MSE and BadPix score confirm our observations. Note that SLFC performs better than our baseline model in terms of multimodal posterior prediction. This clearly shows that the method is indeed able to correctly predict multiple disparity modes. However, the predicted posterior distributions also suffer from poor performance in uncertain regions. Additionally, due to each pixel being optimized separately, the computational cost and therefore runtime of SLFC is several orders of magnitudes higher. One 256px × 256px scene took approximately 18 minutes to compute in parallel on a dual CPU machine with 128 cores, while DPP runs in approximately one second on a single GPU. This makes SLFC, unlike our methods, generally expensive and unsuitable for real-time applications.

We refer to appendix B.2 for an evaluation on the HCI 4D Light Field Dataset [10].

5.4 Conclusion

This chapter investigated the problem of multimodal depth estimation from light fields. We therefore contributed the first light field dataset with multimodal depth ground truth. Additionally, we introduced and compared novel approaches for multimodal light field depth estimation, building on common uncertainty quantification tools. We observe that methods assuming a single valid depth work best if this assumption holds. DPP, which predicts arbitrary posterior distributions, works best in multimodal areas. Our ESE method does not achieve the same performance, but estimates accurate confidence measures even for wide-baseline light fields. We hope that our insights lay the foundations for a new line of depth estimation research that overcomes some long-standing limitations of the field.

6. Outlook and Conclusion

This chapter concludes the thesis by proposing future research directions and summarizing key findings.

6.1 Outlook

First, let us outline potential directions for further research.

Our work is primarily centered on the practicality and applicability of light field depth estimation. While we incorporated a U-Net component as detailed in chapter 4, our exploration of diverse neural network architectures was limited, largely basing our experiments on EPINET [20]. Yet, we believe substantial potential exists for enhancing neural network designs. Given that EPI analysis is a less complex task compared to, for example, image classification, it's plausible that architectures with fewer parameters could achieve comparable depth estimation performance to current leading methods. This would offer multiple benefits: Firstly, a reduction in parameters could lead to improved generalization capability. Secondly, it may result in quicker processing times, potentially enabling real-time application of methods like DPP and broadening the usability of techniques like ESE. Lastly, it could allow for the utilization of more views of the input light field without encountering memory constraints. Present approaches often use only portions of the light field, like the cross or star configurations, due to limited GPU memory. Reducing the parameter count would save memory, thereby enabling the use of more input views, which could lead to more precise estimates, particularly in areas with occlusions.

Moreover, the integration of our methods into specific applications should be investigated further. An obvious application is the shape reconstruction of objects or scenes using a light field camera array. The object could be captured from multiple perspectives, with depth estimated at each angle using methods such as ESE. Subsequently, differentiable rendering could be employed to create a mesh model of the object. In this scenario, all contributions of this thesis would be beneficial: A high-resolution camera array with a wide-baseline would enhance accuracy but would require the use of a method capable of handling its baseline, like EPI-Shift or ESE. Additionally, providing a depth posterior would offer more utility than a single depth value in two scenarios: Firstly, for objects with semi-transparent parts, multiple depth modes could more accurately guide differentiable rendering. Secondly, in uncertain areas, a reliable confidence measure would enable the usage of more sophisticated prior knowledge or regularization strategies to fill these gaps, as opposed to relying on simple smoothness terms commonly used in depth reconstruction.

Exploring integration into other areas such as guided data annotation for semantic segmentation or optical flow could also prove valuable.

6.2 Conclusion

To summarize, this theses extended the practicality of state-of-the-art methods for real-world applications, particularly in the context of wide-baseline camera arrays, semi-transparent and reflective surfaces, and uncertainty quantification.

In chapter 4, we introduced a new learning-based approach for depth estimation from wide-baseline light field recordings. Utilizing EPI-Shift, this method allows for sub-pixel accurate disparity estimation from a wide disparity range, even when trained solely on scenes with a small disparity range. Our framework performs joint classification of integer disparities and regression of disparity-offsets. We employ a U-Net architecture, which has a lower number of parameters relative to the size of its receptive field. This improves the generalization capability, reduces artifacts, and enhances smoothness. The method demonstrates state-of-the-art performance on a publicly available light field benchmark and extends the applicability of narrow-baseline optimized techniques to wide-baseline camera arrays. This adaptation makes our method applicable to light fields recorded with wide-baseline high-resolution camera arrays used in industrial applications and for dataset recording.

In chapter 5, we addressed the problem of multimodal depth posterior estimation from light fields. Posterior regression is helpful for two aspects in the context of light field depth estimation: Firstly, a accurate depth posterior distribution serves as an uncertainty measure which is helpful for downstream applications that depend on the trustworthiness of depth estimates. Secondly, at semi-transparent surfaces and reflections, there exists more than a single depth mode, because objects at multiple depths are visible. We introduced and validated novel approaches for multimodal light field depth estimation, such as UPR, ESE, and DPP. To train and validate our new methods, we introduced the first light field dataset with multimodal depth ground truth. Our findings indicate that methods assuming a single valid depth are most effective when this assumption holds, while DPP performed best in multimodal scenarios. Although ESE did not achieve the same level of performance, it provides accurate confidence measures for wide-baseline light fields. These insights are crucial for advancing depth estimation in complex scenarios and integrating light field depth estimation into applications where knowing the reliability of the estimated depth is key.

The methods introduced in chapter 4 and chapter 5 can be used in various configurations, depending on the desired application and requirements: For wide-baseline recordings, ESE should be used. Alternatively, when uncertainty is not needed for the application, our plain EPI-Shift framework can be used as well. For narrow-baseline recordings, e.g., from plenoptic cameras, DPP should be used to get the best multimodal performance. Alternatively, if multimodal estimation is not needed, but accurate uncertainty is key, UPR is the best choice. Hence, the framework introduced in this thesis covers a broad range of applications.

In conclusion, this thesis has enhanced the methods for light field depth estimation, addressing the limitations of recent research and broadening their applicability to more diverse and challenging real-world applications. We hope that our insights guide future applications and help the research community to overcome some long-standing limitations of the field.

A. Wide-Baseline Light Field Depth Estimation with EPI-Shift

A.1 Additional Experiments

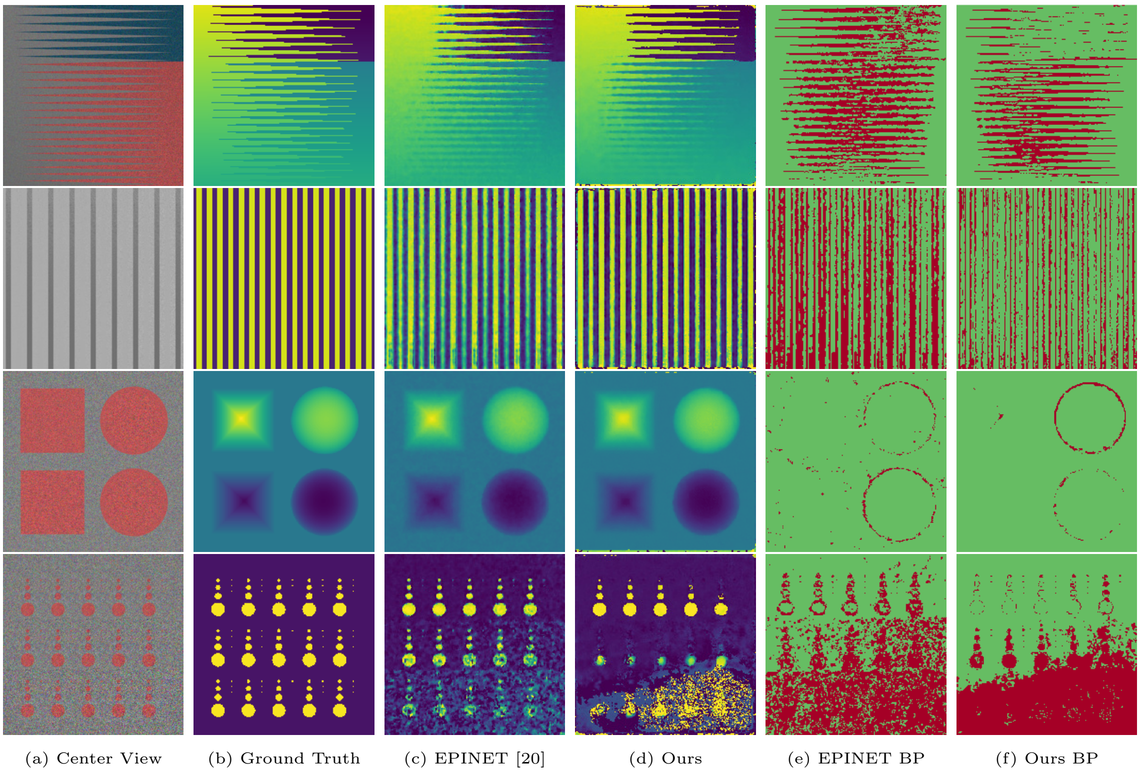


Figure A.1: Results on *stratified* scenes of the HCI 4D Light Field Benchmark [10]. The BadPix score in (e) and (f) shows all pixels (red) exceeding an $\mathcal{L}_1$-distance of 0.07 to the ground truth. Note the improved smoothness on flat surfaces due to the U-Net architecture (compare Scene 3, background). Also note the failure case of our method in Scene 4, caused by strong noise occuring only in the bottom of the image. In those cases, EPI-Shift causes misclassifications, leading to stronger artifacts than EPINET [20].

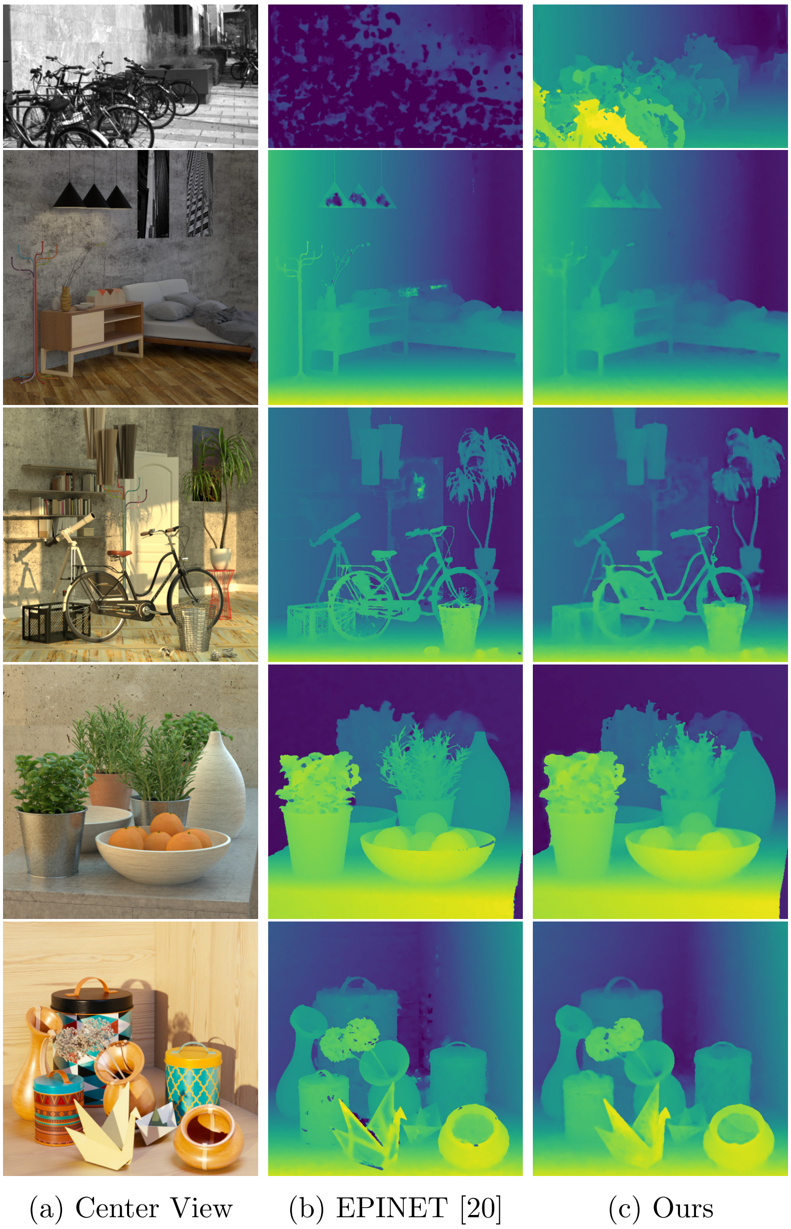


Figure A.2: Results on a second real scene (top) and four additional benchmark scenes [10] without publicly available ground truth. The real recording (top) shows the limitation of EPINET [20] to the disparity range of the training data. Additional benchmark scenes show an improvement in non-textured areas and at extreme disparities (compare Scene 4 (the last scene), background) but also slightly more blurry results of our method. However, blur only occurs within the small regression intervals if two objects are part of the same depth label. At extreme disparities (compare Scene 3, background), our method also performs better due to the hard transitions between adjacent labels.

B. Towards Multimodal Depth Estimation from Light Fields

B.1 Network Architectures

Method	Parameters
EPINET	4612166
UPR	4613300
ESE	4613300
DPP	4778872

Table B.1: Number of trainable parameters for different models.

This section describes the architectures of UPR, ESE and DPP in more detail. The architecture for all of our models is based on EPINET [20]. We input four light field view stacks: horizontal, vertical and two diagonals (vertical stack is visualized in fig. 2.14). Each stack is processed by a separate input stream network. The horizontal and vertical stacks behave similarly when one is rotated by 90°. Therefore we effectively share the weights between those two input streams by applying this rotation to the vertical input and revert it before concatenation. Analogously, we also share weights between the two diagonal input streams. Subsequently, we concatenate the inferred features, and feed them to an output stream. All models and streams share the same basic building block which consists of two convolutions with a kernel size of 2×2. We use an alternating padding of one and zero and a stride of one to maintain the image dimensions. In addition, we apply a ReLU non-linearity after the first convolution and a BN as well as a ReLU layer after the second convolution. Table B.1 shows the total number of trainable parameters for each model. A small difference between the four methods is caused by the variable number of output channels. In the following sections, we describe details, specific to one of the architectures.

All four methods, share the same back bone network. The only differences are the variable number of output channels and one additional output ReLU-layer for DPP. Table B.2 shows the detailed architecture for one input stream. This subnetwork infers features from one light field stack containing nine images with three color channels, thus a total number of $9 \times 3 = 27$ input channels. Each input stream consists of three basic blocks. Because the architecture is based on EPINET [20], we chose the same number of 70 output channels. The features of all input channels are concatenated to a total number of $4 * 70 = 280$ feature channels and fed to the

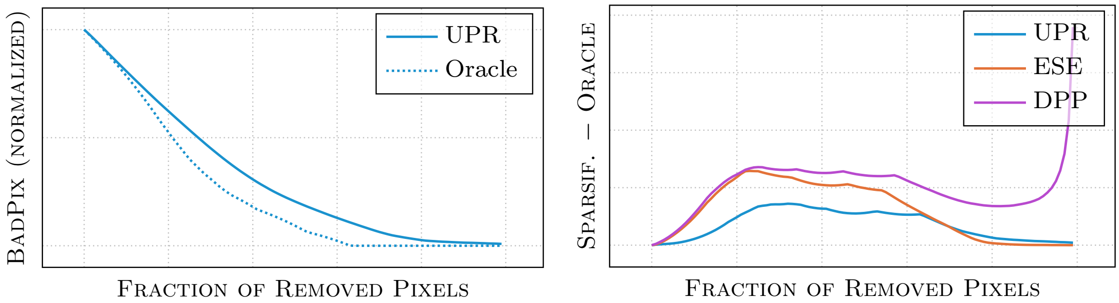
Layer	**Output Size**
LF Stack	$B \times 27 \times H \times W$
2×2 Conv	$B \times 70 \times H \times W$
ReLU	
2×2 Conv	$B \times 70 \times H \times W$
BN	
ReLU	
Repeat Block (2×)	

Table B.2: Input stream of EPINET, UPR, ESE and DPP.

Layer	**Output Size**
Concatenate	$B \times 280 \times H \times W$
2×2 Conv	$B \times 280 \times H \times W$
ReLU	
2×2 Conv	$B \times 280 \times H \times W$
BN	
ReLU	
Repeat Block (6×)	
2×2 Conv	$B \times C_{\text{out}} \times H \times W$
ReLU	
2×2 Conv	$B \times C_{\text{out}} \times H \times W$
(ReLU)	

Table B.3: Output stream of EPINET, UPR, ESE and DPP.

output stream which is illustrated in table B.3. The feed-forward output stream consists of a total number of eight blocks. Both convolutional layers for each block, except the last, output 280 output channels. The last block a certain number of channels, depending on the specific model. Our baseline outputs only one channel, because it directly predicts the disparity for each pixel. For Laplacian distribution prediction, we added a second output channel to also predict the scaling factor b, therefore UPR and ESE have two output channels. The number of discrete disparity 'classes', predicted by DPP, can be chosen arbitrarily. Specifically, we chose 108 output channels, thus 108 'classes', motivated by the common BadPix007 metric.



(a) Sparsification curve

(b) Sparsification Error for all methods

Figure B.1: Unimodal uncertainty quantification on HCI 4D Light Field Dataset. Sparsification results of analyzed methods with respect to the disparity BadPix007.

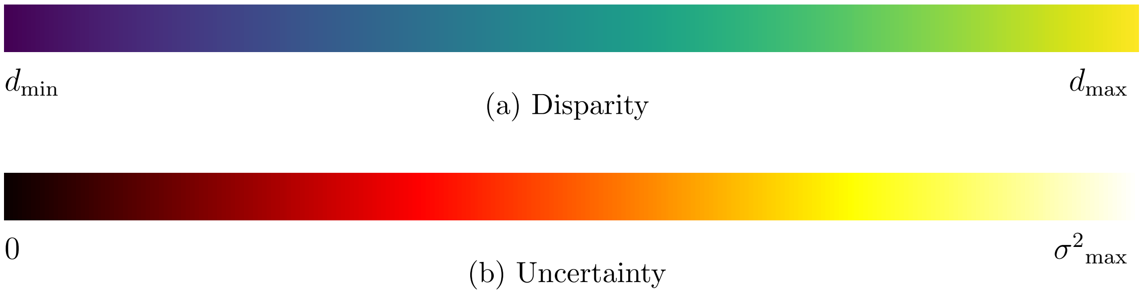
Method	Unimodal Metrics		AuSE $\downarrow$	Time $\downarrow$
	MSE $\downarrow$	BadPix $\downarrow$		(in sec)
BASE	**0.011**	0.065	-	**0.480**
UPR	0.012	0.056	**0.060**	0.481
ESE	0.163	0.088	0.091	14.863
DPP	0.018	**0.044**	0.110	0.783

Table B.4: Evaluation on HCI dataset [10], from left to right: MSE and the common BadPix007 score (percentage of pixels with $|d - d_{\text{true}}| > 0.07$), AuSE, runtime of one forward pass. Lower is better.

B.2 Additional Experiments

This sections shows the additional evaluation of our methods on the commonly used HCI 4D Light Field Dataset [10]. Like previous methods [20], we used the 16 *additional* scenes as our training dataset and the four *training* scenes for validation. As this dataset only contains a single ground truth depth, we used the unimodal loss functions $\mathcal{L}_x$. All other training parameters remain the same as mentioned in chapter 5.

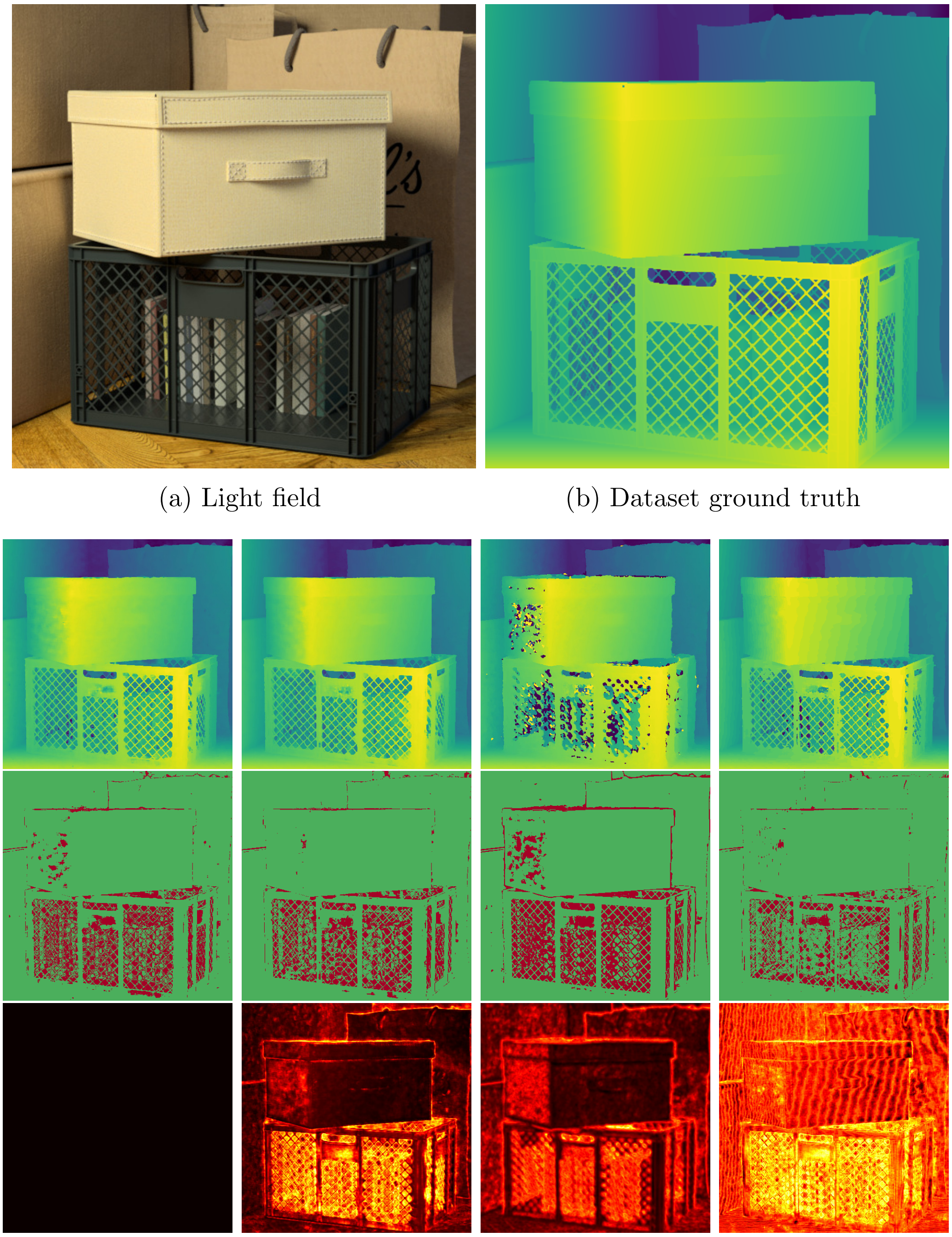
Figure B.1 and table B.4 show our experimental results, which are overall very consistent with the experiments on our randomly generated multimodal dataset. DPP performs best with respect to the amount of accurately predicted pixels (BadPix) but is overconfident which is clearly visible in the sparsification error. In contrast, UPR and ESE deliver a better sparsification performance. Qualitative results are shown in fig. B.3 to fig. B.6.



(a) Disparity

(b) Uncertainty

Figure B.2: Color maps used for results. Disparity and uncertainty maps are normalized to enhance visibility.



(a) Light field

(b) Dataset ground truth

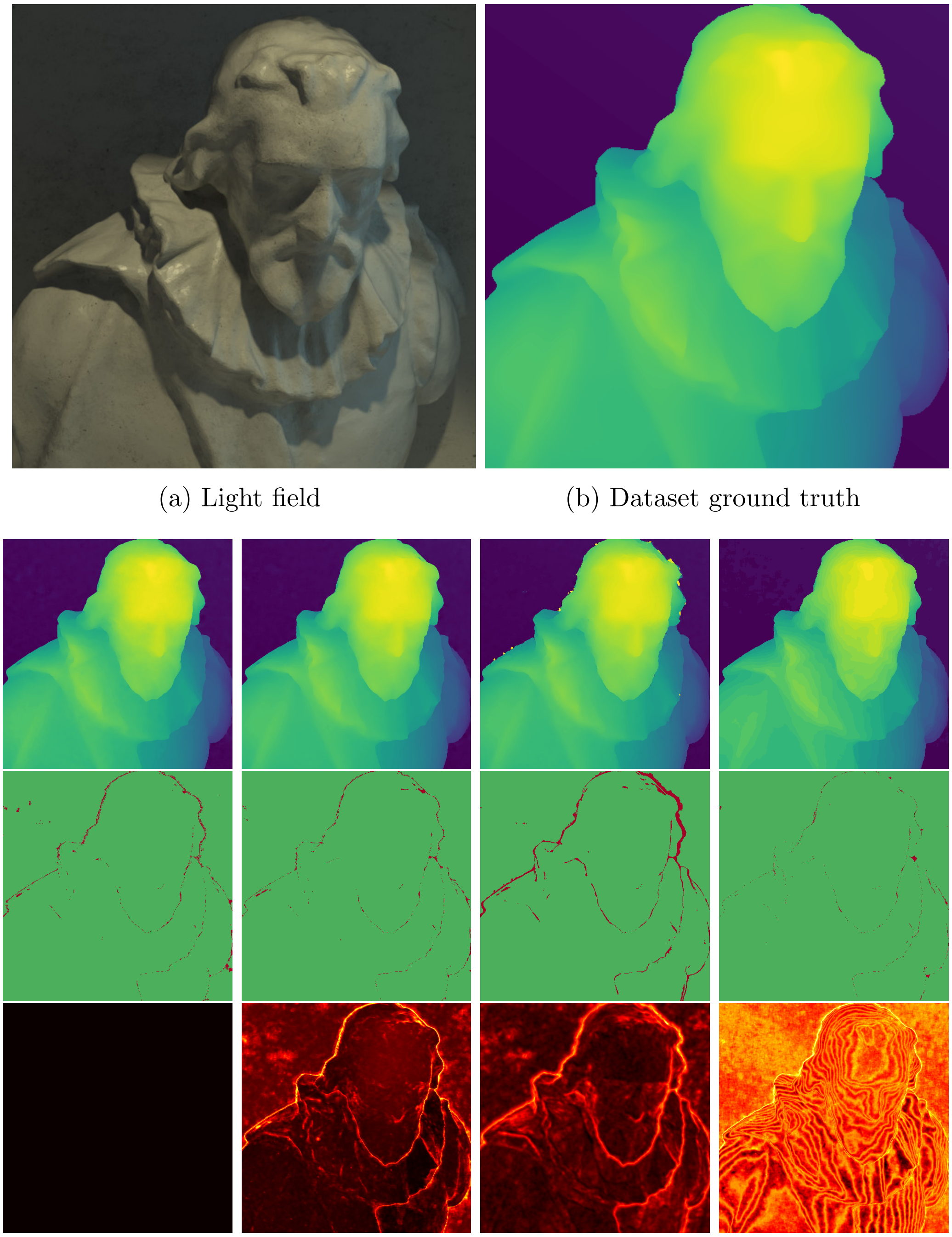
(c) BASE

(d) UPR

(e) ESE

(f) DPP

Figure B.3: Results of the four posterior prediction methods ((c) - (f)) for ‘boxes’ scene. Top: output disparity (most likely mode). Center: per-pixel BadPix metric (a pixel i is red if $|d - d_{\text{true}}| > 0.07$). Bottom: per-pixel uncertainty σ^2 (non-existent for baseline method).



(a) Light field

(b) Dataset ground truth

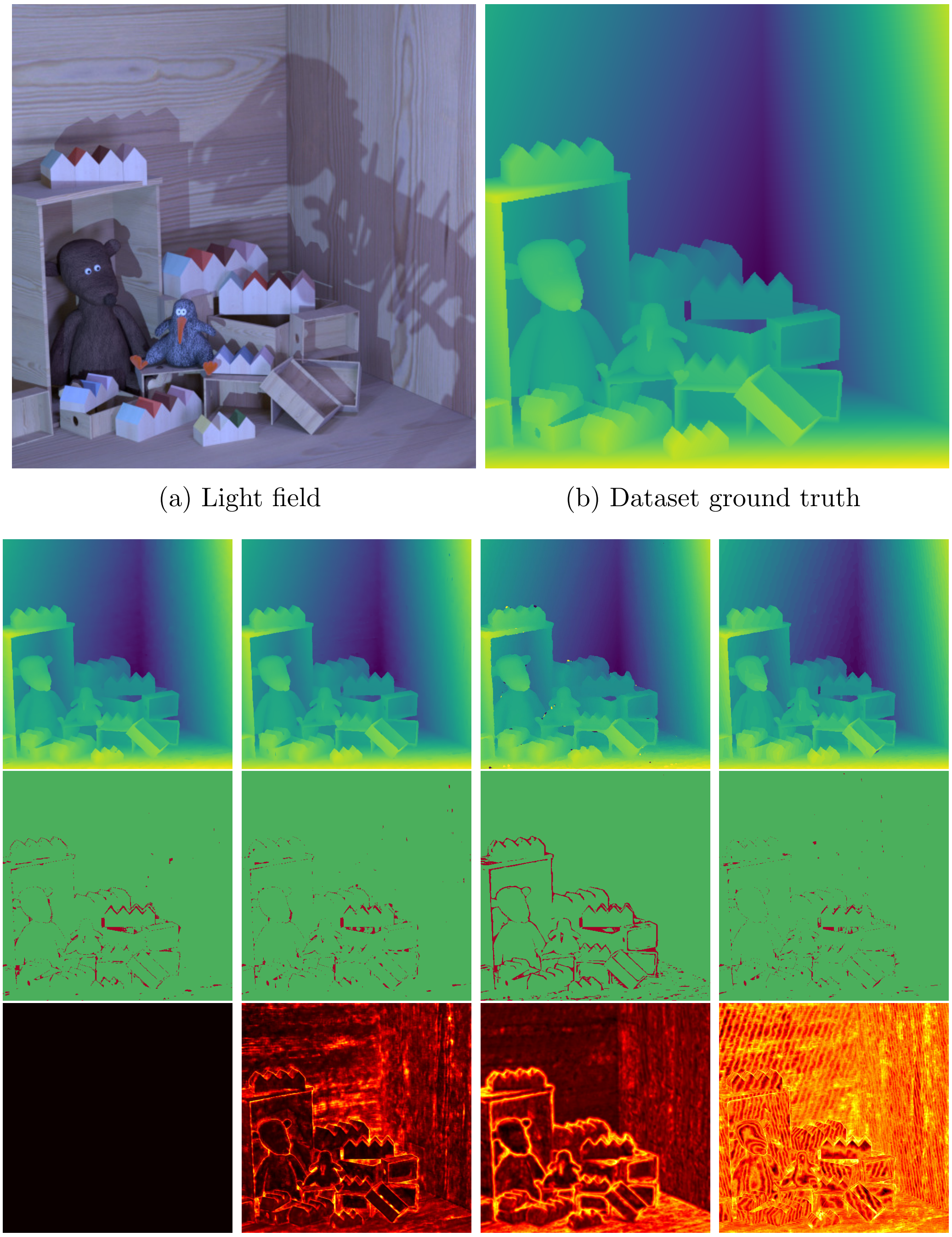
(c) BASE

(d) UPR

(e) ESE

(f) DPP

Figure B.4: Results of the four posterior prediction methods ((c) - (f)) for ‘cotton’ scene. Top: output disparity (most likely mode). Center: per-pixel BadPix metric (a pixel i is red if $|d - d_{\text{true}}| > 0.07$). Bottom: per-pixel uncertainty σ^2 (non-existent for baseline method).



(a) Light field

(b) Dataset ground truth

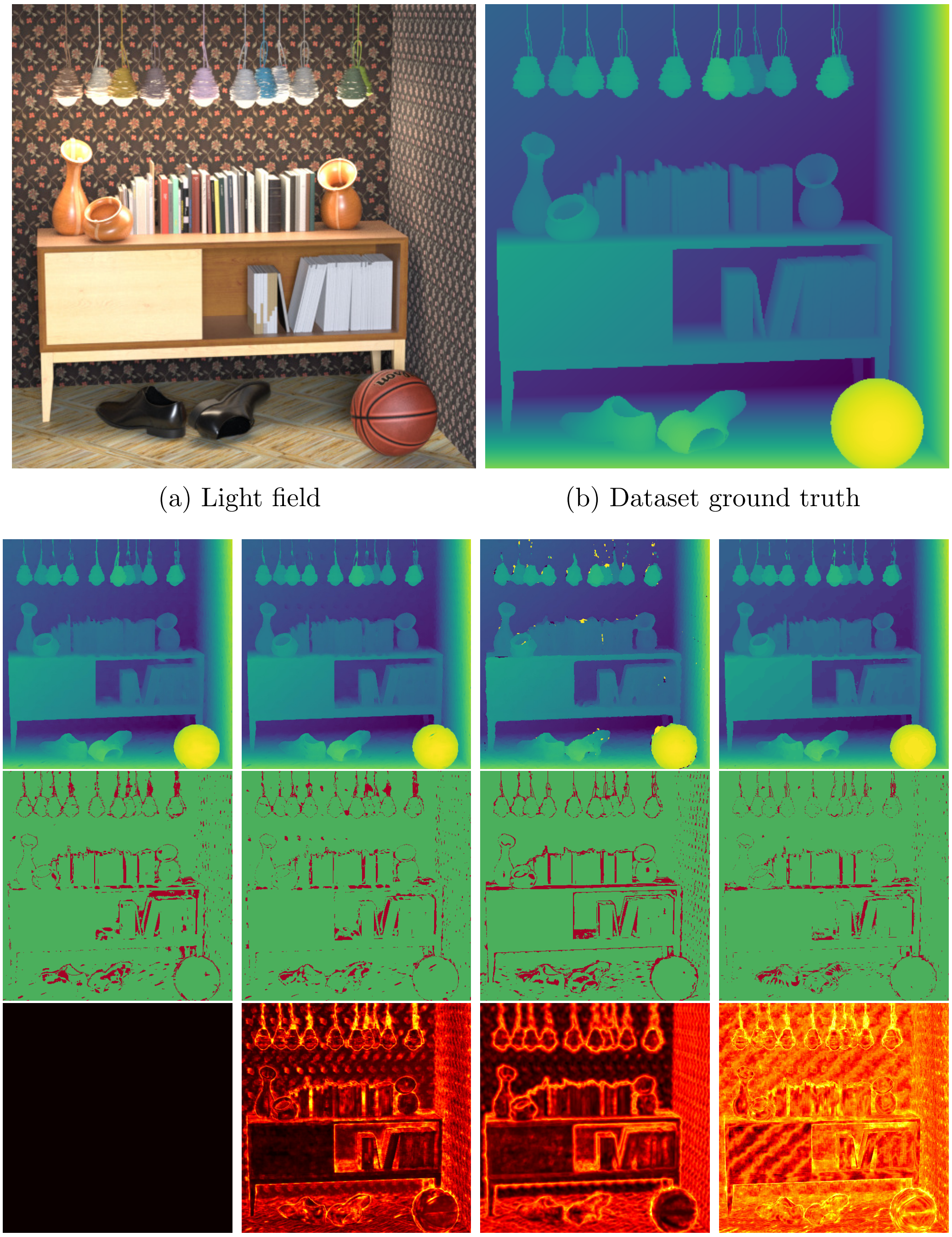
(c) BASE

(d) UPR

(e) ESE

(f) DPP

Figure B.5: Results of the four posterior prediction methods ((c) - (f)) for 'dino' scene. Top: output disparity (most likely mode). Center: per-pixel BadPix metric (a pixel i is red if $|d - d_{\text{true}}| > 0.07$). Bottom: per-pixel uncertainty σ^2 (non-existent for baseline method).



(a) Light field

(b) Dataset ground truth

(c) BASE

(d) UPR

(e) ESE

(f) DPP

Figure B.6: Results of the four posterior prediction methods ((c) - (f)) for ‘sideboard’ scene. Top: output disparity (most likely mode). Center: per-pixel BadPix metric (a pixel i is red if $|d - d_{\text{true}}| > 0.07$). Bottom: per-pixel uncertainty σ^2 (non-existent for baseline method).

List of Tables

List of Figures

Bibliography

[1] H. Schilling, M. Gutsche, A. Brock, D. Spath, C. Rother, and K. Krispin, "Mind the gap-a benchmark for dense depth prediction beyond lidar", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 338–339. DOI: https://doi.org/10.1109/CVPRW50498.2020.00177.

[2] T. Wang, H. Sheng, R. Chen, *et al.*, "Light field depth estimation: A comprehensive survey from principles to future", *High-Confidence Computing*, p. 100 187, 2023. DOI: https://doi.org/10.1016/j.hcc.2023.100187.

[3] M. Levoy and P. Hanrahan, "Light field rendering", in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 31–42. DOI: https://doi.org/10.1145/3596711.3596759.

[4] N. Joshi, *Uscd/merl light field repository*, UCSD Vision and Graphics Laboratories and Mitsubishi Electric Research Laboratory, 2007. [Online]. Available: http://neelj.com/data/lfarchive/.

[5] V. Vaish and A. Adams, *The (new) stanford light field archive*, Stanford Computer Graphics Laboratory, 2008. [Online]. Available: http://lightfield.stanford.edu/.

[6] S. Wanner, S. Meister, and B. Goldlücke, "Datasets and benchmarks for densely sampled 4d light fields", in *International Symposium on Vision, Modeling, and Visualization*, 2013. DOI: https://doi.org/10.2312/PE.VMV.VMV13.225-226.

[7] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera", Ph.D. dissertation, Stanford University, 2005.

[8] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras", in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 673–680. DOI: https://doi.org/10.1109/ICCV.2013.89.

[9] H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry", in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3451–3459. DOI: https://doi.org/10.1109/ICCV.2015.394.

[10] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields", in *Asian Conference on Computer Vision*, Springer, 2016, pp. 19–34. DOI: https://doi.org/10.1007/978-3-319-54187-7_2.

[11] M. Feng, Y. Wang, J. Liu, L. Zhang, H. F. Zaki, and A. Mian, “Benchmark data set and method for depth estimation from light field images”, *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3586–3598, 2018. DOI: `https://doi.org/10.1109/TIP.2018.2814217`.

[12] H. Sheng, R. Cong, D. Yang, R. Chen, S. Wang, and Z. Cui, “Urbanlf: A comprehensive light field dataset for semantic segmentation of urban scenes”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7880–7893, 2022. DOI: `https://doi.org/10.1109/TCSVT.2022.3187664`.

[13] M. Tomkins, *Well, duh: Lytro finally realizes nobody wants refocusable, low-res, overpriced cameras*, Imaging Resource, 2016. [Online]. Available: `https://www.imaging-resource.com/news/2016/04/05/well-duh-lytro-finally-realizes-nobody-wants-refocusable-low-res-cameras`.

[14] A. Robertson, *Vr camera maker lytro is shutting down, and former employees are going to google*, The Verge, 2018. [Online]. Available: `https://www.theverge.com/2018/3/27/17166038/lytro-light-field-camera-company-shuts-down-google-hiring`.

[15] B. Myers, *Lytro’s demise and the future of light field cameras*, Android Authority, 2018. [Online]. Available: `https://www.androidauthority.com/lytro-light-field-camera-869929/`.

[16] M. Schindler-Kotschka, *Top hit rates in bin picking for benteler automotive*, Case Study, 2021. [Online]. Available: `https://www.hdvisionsystems.com/en/blog/bin-picking-for-benteler/`.

[17] M. Schindler-Kotschka, *Fully automated quality assurance in collaboration with pütz group*, 2021. [Online]. Available: `https://www.hdvisionsystems.com/en/blog/quality-inspection-innovision/`.

[18] J. Winter, *Gripping door hinges: Implementing complex shapes and surfaces in bin picking*, 2021. [Online]. Available: `https://www.hdvisionsystems.com/en/blog/gripping-door-pivots-with-bin-picking/`.

[19] H. Schilling, *Why rabbitai makes a difference*, 2020. [Online]. Available: `https://rabbitai.de/blog/why-rabbitai-makes-a-difference/`.

[20] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, “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. DOI: `https://doi.org/10.1109/CVPR.2018.00499`.

[21] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, “Attention-based view selection networks for light-field disparity estimation”, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 095–12 103. DOI: `https://doi.org/10.1609/aaai.v34i07.6888`.

[22] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, “D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1281–1292. DOI: `https://doi.org/10.1109/CVPR42600.2020.00136`.

[23] E. H. Adelson and J. R. Bergen, “The plenoptic function and the elements of early vision”, *Computational models of visual processing*, vol. 1, no. 2, pp. 3–20, 1991. DOI: https://doi.org/10.7551/mitpress/2002.003.0004.

[24] J. Kepler, *Astronomiae pars optica.* 1939.

[25] I. Newton, *Opticks:: Or, A Treatise of the Reflections, Refractions, Inflections and Colours of Light.* 1704. DOI: https://doi.org/10.1097/00006324-193305000-00006.

[26] M. Faraday, “Thoughts on ray-vibrations”, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 28, no. 188, pp. 345–350, 1846. DOI: https://doi.org/10.1017/CBO9781139383165.018.

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis”, *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. DOI: https://doi.org/10.1145/3503250.

[28] B. Wilburn, N. Joshi, V. Vaish, *et al.*, “High performance imaging using large camera arrays”, in *ACM SIGGRAPH 2005 Papers*, 2005, pp. 765–776. DOI: https://doi.org/10.1145/1186822.1073259.

[29] M. Levoy and J. Shade, *A light field of michelangelo’s statue of night*, 1999. [Online]. Available: https://accademia.stanford.edu/mich/lightfield-of-night/lightfield-of-night.html.

[30] C. Perwass and L. Wietzke, “Single lens 3d-camera with extended depth-of-field”, in *Human vision and electronic imaging XVII*, SPIE, vol. 8291, 2012, pp. 45–59. DOI: https://doi.org/10.1117/12.909882.

[31] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, “Image and depth from a conventional camera with a coded aperture”, *ACM transactions on graphics (TOG)*, vol. 26, no. 3, 70–es, 2007. DOI: https://doi.org/10.1145/1275808.1276464.

[32] G. Lippmann, *Photography. - reversible prints. integral photographs*, Note, 1908. [Online]. Available: https://people.csail.mit.edu/fredo/PUBLI/Lippmann.pdf.

[33] E. H. Adelson and J. Y. Wang, “Single lens stereo with a plenoptic camera”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 99–106, 1992. DOI: https://doi.org/10.1109/34.121783.

[34] I. Ihm, S. Park, and R. K. Lee, “Rendering of spherical light fields”, in *Proceedings The Fifth Pacific Conference on Computer Graphics and Applications*, IEEE, 1997, pp. 59–68. DOI: https://doi.org/10.1109/PCCGA.1997.626172.

[35] H.-Y. Shum and L.-W. He, “Rendering with concentric mosaics”, in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 299–306. DOI: https://doi.org/10.1145/311535.311573.

[36] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, “A system for acquiring, processing, and rendering panoramic light field stills for virtual reality”, *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018. DOI: https://doi.org/10.1145/3272127.3275031.

[37] M. Broxton, J. Flynn, R. Overbeck, *et al.*, "Immersive light field video with a layered mesh representation", *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 86–1, 2020. DOI: `https://doi.org/10.1145/3386569.3392485`.

[38] I. Failes, *VFX Artifacts: The Bullet Time rig from 'The Matrix'*, Befores and Afters, 2021. [Online]. Available: `https://beforesandafters.com/2021/07/15/vfx-artifacts-the-bullet-time-rig-from-the-matrix/`.

[39] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array", in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, IEEE, vol. 2, 2004, pp. II–II. DOI: `https://doi.org/10.1109/CVPR.2004.1315176`.

[40] J. An, S. Park, and I. Ihm, "Construction of a flexible and scalable 4d light field camera array using raspberry pi clusters", *The Visual Computer*, vol. 35, pp. 1475–1488, 2019. DOI: `https://doi.org/10.1007/s00371-018-1512-z`.

[41] T. Kanade, H. Saito, and S. Vedula, *The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams.* Citeseer, 1998.

[42] B. S. Wilburn, M. Smulski, H.-H. K. Lee, and M. A. Horowitz, "The light field video camera", in *Media Processors 2002*, SPIE, vol. 4674, 2001, pp. 29–36. DOI: `https://doi.org/10.1117/12.451074`.

[43] Raspberry Pi Foundation, *Raspberry pi.* [Online]. Available: `https://www.raspberrypi.org/`.

[44] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion", *International journal of computer vision*, vol. 1, no. 1, pp. 7–55, 1987. DOI: `https://doi.org/10.1007/BF00128525`.

[45] The LEGO Group, *Lego mindstorms.* [Online]. Available: `https://www.lego.com/en-us/themes/mindstorms/about`.

[46] M. Levoy, K. Pulli, B. Curless, *et al.*, "The digital michelangelo project: 3d scanning of large statues", in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 131–144. DOI: `https://doi.org/10.1145/344779.344849`.

[47] C. Hahne and A. Aggoun, "Plenopticam v1. 0: A light-field imaging framework", *IEEE Transactions on Image Processing*, vol. 30, pp. 6757–6771, 2021. DOI: `https://doi.org/10.1109/TIP.2021.3095671`.

[48] L. Wietzke, *3d camera applications*, 2016. [Online]. Available: `https://raytrix.de/wp-content/uploads/software/Life-Science-Raytrix-3D-Light-Field-Camera.pdf`.

[49] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, "Programmable aperture photography: Multiplexed light field acquisition", in *ACM siggraph 2008 papers*, 2008, pp. 1–10. DOI: `https://doi.org/10.1145/1399504.1360654`.

[50] R. Dicke, "Scatter-hole cameras for x-rays and gamma rays", *Astrophysical Journal, vol. 153, p. L101*, vol. 153, p. L101, 1968. DOI: `https://doi.org/10.1086/180230`.

[51] J. Ables, "Fourier transform photography: A new method for x-ray astronomy", *Publications of the Astronomical Society of Australia*, vol. 1, no. 4, pp. 172–173, 1968. DOI: https://doi.org/10.1017/S1323358000011292.

[52] E. E. Fenimore and T. M. Cannon, "Coded aperture imaging with uniformly redundant arrays", *Applied optics*, vol. 17, no. 3, pp. 337–347, 1978. DOI: https://doi.org/10.1364/AO.17.000337.

[53] J. Liu, C. Zaouter, X. Liu, S. A. Patten, and J. Liang, "Coded-aperture broadband light field imaging using digital micromirror devices", *Optica*, vol. 8, no. 2, pp. 139–142, 2021. DOI: https://doi.org/10.1364/OPTICA.413938.

[54] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, "Learning to capture light fields through a coded aperture camera", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434. DOI: https://doi.org/10.1007/978-3-030-01234-2_26.

[55] G. Wetzstein, *Synthetic light field archive*, 2013. [Online]. Available: https://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php.

[56] Y. Li, Q. Wang, L. Zhang, and G. Lafruit, "A lightweight depth estimation network for wide-baseline light fields", *IEEE Transactions on Image Processing*, vol. 30, pp. 2288–2300, 2021. DOI: https://doi.org/10.1109/TIP.2021.3051761.

[57] J. L. Posdamer and M. D. Altschuler, "Surface measurement by space-encoded projected beam systems", *Computer graphics and image processing*, vol. 18, no. 1, pp. 1–17, 1982. DOI: https://doi.org/10.1016/0146-664X(82)90096-X.

[58] D. Nitzan, A. E. Brain, and R. O. Duda, "The measurement and use of registered reflectance and range data in scene analysis", *Proceedings of the IEEE*, vol. 65, no. 2, pp. 206–220, 1977. DOI: https://doi.org/10.1109/PROC.1977.10458.

[59] G. Smith, "The early laser years at hughes aircraft company", *IEEE journal of quantum electronics*, vol. 20, no. 6, pp. 577–584, 1984. DOI: https://doi.org/10.1109/JQE.1984.1072445.

[60] A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images", *Advances in neural information processing systems*, vol. 18, 2005.

[61] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review", *Neurocomputing*, vol. 438, pp. 14–33, 2021. DOI: https://doi.org/10.1016/j.neucom.2020.12.089.

[62] S. Ullman, "The interpretation of structure from motion", *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979. DOI: https://doi.org/10.7551/mitpress/3877.003.0009.

[63] A. P. Pentland, "A new sense for depth of field", *IEEE transactions on pattern analysis and machine intelligence*, no. 4, pp. 523–531, 1987. DOI: https://doi.org/10.1109/TPAMI.1987.4767940.

[64] D. Marr, T. Poggio, E. C. Hildreth, and W. E. L. Grimson, *A computational theory of human stereo vision*. Springer, 1991. DOI: https://doi.org/10.1007/978-1-4684-6775-8_11.

[65] T. Schops, J. L. Schonberger, S. Galliani, *et al.*, “A multi-view stereo benchmark with high-resolution images and multi-camera videos”, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3260–3269. DOI: `https://doi.org/10.1109/CVPR.2017.272`.

[66] M. Okutomi and T. Kanade, “A multiple-baseline stereo”, *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 4, pp. 353–363, 1993. DOI: `https://doi.org/10.1109/34.206955`.

[67] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022. DOI: `https://doi.org/10.1007/978-1-84882-935-0`.

[68] A. Fusiello, E. Trucco, and A. Verri, “A compact algorithm for rectification of stereo pairs”, *Machine vision and applications*, vol. 12, pp. 16–22, 2000. DOI: `https://doi.org/10.1007/s001380050120`.

[69] H. Schilling, “Light field camera arrays calibration and depth estimation”, Ph.D. dissertation, 2023.

[70] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, “Extracting layers and analyzing their specular properties using epipolar-plane-image analysis”, *Computer vision and image understanding*, vol. 97, no. 1, pp. 51–85, 2005. DOI: `https://doi.org/10.1016/j.cviu.2004.06.001`.

[71] H. Schilling, M. Diebold, C. Rother, and B. Jähne, “Trust your model: Light field depth estimation with inline occlusion handling”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4530–4538. DOI: `https://doi.org/10.1109/CVPR.2018.00476`.

[72] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, “Robust depth estimation for light field via spinning parallelogram operator”, *Computer Vision and Image Understanding*, vol. 145, pp. 148–159, 2016. DOI: `https://doi.org/10.1016/j.cviu.2015.12.007`.

[73] R. T. Collins, “A space-sweep approach to true multi-image matching”, in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 1996, pp. 358–363. DOI: `https://doi.org/10.1109/CVPR.1996.517097`.

[74] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, *et al.*, “Local light field fusion: Practical view synthesis with prescriptive sampling guidelines”, *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019. DOI: `https://doi.org/10.1145/3306346.3322980`.

[75] Y. Y. Schechner and N. Kiryati, “Depth from defocus vs. stereo: How different really are they?”, *International Journal of Computer Vision*, vol. 39, no. 2, pp. 141–162, 2000. DOI: `https://doi.org/10.1109/ICPR.1998.712074`.

[76] S. Baker, R. Szeliski, and P. Anandan, “A layered approach to stereo reconstruction”, in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*, IEEE, 1998, pp. 434–441. DOI: `https://doi.org/10.1109/CVPR.1998.698642`.

[77] M. Bleyer, C. Rhemann, and C. Rother, “Patchmatch stereo-stereo matching with slanted support windows.”, in *Bmvc*, vol. 11, 2011, pp. 1–11. DOI: `https://doi.org/10.5244/C.25.14`.

[78] M. Matoušek, T. Werner, and V. Hlavác, "Accurate correspondences from epipolar plane images", in *Proc. Computer Vision Winter Workshop*, Citeseer, 2001, pp. 181–189.

[79] R. Bellman, "Dynamic programming", *science*, vol. 153, no. 3731, pp. 34–37, 1966. DOI: `https://doi.org/10.1126/science.153.3731.34`.

[80] H.-G. Jeon, J. Park, G. Choe, *et al.*, "Accurate depth map estimation from a lenslet light field camera", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1547–1555. DOI: `https://doi.org/10.1109/CVPR.2015.7298762`.

[81] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 51, no. 2, pp. 271–279, 1989. DOI: `https://doi.org/10.1111/j.2517-6161.1989.tb01764.x`.

[82] T. E. Bishop and P. Favaro, "Plenoptic depth estimation from multiple aliased views", in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, IEEE, 2009, pp. 1622–1629. DOI: `https://doi.org/10.1109/ICCVW.2009.5457420`.

[83] S. Heber and T. Pock, "Shape from light field meets robust pca", in *European Conference on Computer Vision*, Springer, 2014, pp. 751–767. DOI: `https://doi.org/10.1007/978-3-319-10599-4_48`.

[84] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution", *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 606–619, 2013. DOI: `https://doi.org/10.1109/TPAMI.2013.147`.

[85] A. Neri, M. Carli, and F. Battisti, "A multi-resolution approach to depth field estimation in dense image arrays", in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2015, pp. 3358–3362. DOI: `https://doi.org/10.1109/ICIP.2015.7351426`.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, vol. 25, 2012. DOI: `https://doi.org/10.1145/3065386`.

[87] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite", in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 3354–3361. DOI: `https://doi.org/10.1109/CVPR.2012.6248074`.

[88] D. Scharstein, H. Hirschmüller, Y. Kitajima, *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth", in *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, Springer, 2014, pp. 31–42. DOI: `https://doi.org/10.1007/978-3-319-11752-2_3`.

[89] S. Heber and T. Pock, "Convolutional networks for shape from light field", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3746–3754. DOI: `https://doi.org/10.1109/CVPR.2016.407`.

[90] S. Heber, W. Yu, and T. Pock, “U-shaped networks for shape from light field.”, in *BMVC*, vol. 3, 2016, p. 5. DOI: https://doi.org/10.5244/C.30.37.

[91] J. Chen, S. Zhang, and Y. Lin, “Attention-based multi-level fusion network for light field depth estimation”, in *Proc AAAI Conf Artif Intell*, vol. 35, 2021, pp. 1009–1017. DOI: https://doi.org/10.1609/aaai.v35i2.16185.

[92] W. Zhou, E. Zhou, Y. Yan, L. Lin, and A. Lumsdaine, “Learning depth cues from focal stack for light field depth estimation”, in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 1074–1078. DOI: https://doi.org/10.1109/ICIP.2019.8804270.

[93] W. Zhou, X. Wei, Y. Yan, W. Wang, and L. Lin, “A hybrid learning of multimodal cues for light field depth estimation”, *Digital Signal Processing*, vol. 95, p. 102 585, 2019. DOI: https://doi.org/10.1016/j.dsp.2019.102585.

[94] T. Werner, T. Pajdla, V. Hlavác, *et al.*, “Correspondence by tracking edges in a dense sequence for image-based scene representation”, in *Proceedings of the Czech Pattern Recognition Workshop*, Citeseer, vol. 97, 1997, pp. 64–68.

[95] D. Dansereau and L. Bruton, “Gradient-based depth estimation from 4d light fields”, in *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No. 04CH37512)*, IEEE, vol. 3, 2004, pp. III–549. DOI: https://doi.org/10.1109/ISCAS.2004.1328805.

[96] C. Frese and I. Gheţa, “Robust depth estimation by fusion of stereo and focus series acquired with a camera array”, in *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, IEEE, 2006, pp. 243–248. DOI: https://doi.org/10.1109/MFI.2006.265618.

[97] Y.-H. Kao, C.-K. Liang, L.-W. Chang, and H. H. Chen, “Depth detection of light field”, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*, IEEE, vol. 1, 2007, pp. I–893. DOI: https://doi.org/10.1109/ICASSP.2007.366052.

[98] T. E. Bishop and P. Favaro, “Full-resolution depth map estimation from an aliased plenoptic light field”, in *Asian Conference on Computer Vision*, Springer, 2010, pp. 186–200. DOI: https://doi.org/10.1007/978-3-642-19309-5_15.

[99] C.-C. Chen, S.-C. F. Chiang, X.-X. Huang, M.-S. Su, and Y.-C. Lu, “Depth estimation of light field data from pinhole-masked dslr cameras”, in *2010 IEEE International Conference on Image Processing*, IEEE, 2010, pp. 1769–1772. DOI: https://doi.org/10.1109/ICIP.2010.5653375.

[100] T. E. Bishop and P. Favaro, “The light field camera: Extended depth of field, aliasing, and superresolution”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 5, pp. 972–986, 2011. DOI: https://doi.org/10.1109/TPAMI.2011.168.

[101] P. T. Kovacs and F. Zilly, “3d capturing using multi-camera rigs, real-time depth estimation and depth-based content creation for multi-view and light-field auto-stereoscopic displays”, in *ACM SIGGRAPH 2012 Emerging Technologies*, 2012, pp. 1–1. DOI: https://doi.org/10.1145/2343456.2343457.

[102] S. Wanner, J. Fehr, and B. Jähne, "Generating epi representations of 4d light fields with a single lens focused plenoptic camera", in *International Symposium on Visual Computing*, Springer, 2011, pp. 90–101. DOI: https://doi.org/10.1007/978-3-642-24028-7_9.

[103] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields", in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 41–48. DOI: https://doi.org/10.1109/CVPR.2012.6247656.

[104] S. Heber, R. Ranftl, and T. Pock, "Variational shape from light field", in *Energy Minimization Methods in Computer Vision and Pattern Recognition: 9th International Conference, EMMCVPR 2013, Lund, Sweden, August 19-21, 2013. Proceedings 9*, Springer, 2013, pp. 66–79. DOI: https://doi.org/10.1007/978-3-642-40395-8_6.

[105] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections", *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 1–12, 2013. DOI: https://doi.org/10.1145/2461912.2461914.

[106] Z. Yu, X. Guo, H. Lin, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching", in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2792–2799. DOI: https://doi.org/10.1109/ICCV.2013.347.

[107] C.-W. Chang, M.-R. Chen, P.-H. Hsu, and Y.-C. Lu, "A pixel-based depth estimation algorithm and its hardware implementation for 4-d light field data", in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2014, pp. 786–789. DOI: https://doi.org/10.1109/ISCAS.2014.6865253.

[108] C. Chen, H. Lin, Z. Yu, S. Bing Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1518–1525. DOI: https://doi.org/10.1109/CVPR.2014.197.

[109] M. W. Tao, T.-C. Wang, J. Malik, and R. Ramamoorthi, "Depth estimation for glossy surfaces with light-field cameras", in *European Conference on Computer Vision*, Springer, 2014, pp. 533–547. DOI: https://doi.org/10.1007/978-3-319-16181-5_41.

[110] I. Tosic and K. Berkner, "Light field scale-depth space transform for dense depth estimation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 435–442. DOI: https://doi.org/10.1109/CVPRW.2014.71.

[111] H.-H. Chen, C.-T. Huang, S.-S. Wu, C.-L. Hung, T.-C. Ma, and L.-G. Chen, "A 1920× 1080 30fps 611 mw five-view depth-estimation processor for light-field applications", in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, IEEE, 2015, pp. 1–3. DOI: https://doi.org/10.1109/ISSCC.2015.7063106.

[112] M. Diebold, O. Blum, M. Gutsche, *et al.*, “Light-field camera design for high-accuracy depth estimation”, in *Videometrics, Range Imaging, and Applications XIII*, International Society for Optics and Photonics, vol. 9528, 2015, p. 952 803. DOI: https://doi.org/10.1117/12.2184845.

[113] S. Im, H.-G. Jeon, H. Ha, and I. S. Kweon, “Depth estimation from light field cameras”, in *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, IEEE, 2015, pp. 190–191. DOI: https://doi.org/10.1109/URAI.2015.7358863.

[114] J. Li, M. Lu, and Z.-N. Li, “Continuous depth map reconstruction from light fields”, *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3257–3265, 2015. DOI: https://doi.org/10.1109/ICME.2013.6607557.

[115] H. Lv, K. Gu, Y. Zhang, and Q. Dai, “Light field depth estimation exploiting linear structure in epi”, in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2015, pp. 1–6. DOI: https://doi.org/10.1109/ICMEW.2015.7169836.

[116] R. J. Marshall, C. J. Meah, M. Turola, *et al.*, “Improving depth estimation from a plenoptic camera by patterned illumination”, in *Videometrics, Range Imaging, and Applications XIII*, SPIE, vol. 9528, 2015, pp. 365–370. DOI: https://doi.org/10.1117/12.2184742.

[117] H. Sheng, S. Zhang, G. Zhu, and Z. Xiong, “Guided integral filter for light field stereo matching”, in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2015, pp. 852–856. DOI: https://doi.org/10.1109/ICIP.2015.7350920.

[118] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, “Depth from shading, defocus, and correspondence using light-field angular coherence”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1940–1948. DOI: https://doi.org/10.1109/CVPR.2015.7298804.

[119] M. W. Tao, *Unified Multi-Cue Depth Estimation from Light-Field Images: Correspondence, Defocus, Shading, and Specularity.* University of California, Berkeley, 2015.

[120] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, “Occlusion-aware depth estimation using light-field cameras”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3487–3495. DOI: https://doi.org/10.1109/ICCV.2015.398.

[121] D. Antensteiner, S. Štolc, and R. Huber-Mörk, “Depth estimation with light field and photometric stereo data using energy minimization”, in *Iberoamerican Congress on Pattern Recognition*, Springer, 2016, pp. 175–183. DOI: https://doi.org/10.1007/978-3-319-52277-7_22.

[122] O. Johannsen, A. Sulc, and B. Goldluecke, “Occlusion-aware depth estimation using sparse light field coding”, in *German Conference on Pattern Recognition*, Springer, 2016, pp. 207–218. DOI: https://doi.org/10.1007/978-3-319-45886-1_17.

[123] O. Johannsen, A. Sulc, and B. Goldluecke, “What sparse light field coding reveals about scene structure”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3262–3270. DOI: https://doi.org/10.1109/CVPR.2016.355.

[124] M. G. S. Prabha, “Continuous depth map reconstruction from light fields”, *Journal for Studies in Management and Planning*, vol. 2, no. 12, pp. 92–101, 2016. DOI: https://doi.org/10.1109/TIP.2015.2440760.

[125] M. S. Sajjadi, R. Köhler, B. Schölkopf, and M. Hirsch, “Depth estimation through a generative model of light field synthesis”, in *German Conference on Pattern Recognition*, Springer, 2016, pp. 426–438. DOI: https://doi.org/10.1007/978-3-319-45886-1_35.

[126] X. Sun, Z. Xu, N. Meng, E. Y. Lam, and H. K.-H. So, “Data-driven light field depth estimation using deep convolutional neural networks”, in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2016, pp. 367–374. DOI: https://doi.org/10.1109/IJCNN.2016.7727222.

[127] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, “Depth estimation with occlusion modeling using light-field cameras”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2170–2181, 2016. DOI: https://doi.org/10.1109/TPAMI.2016.2515615.

[128] W. Williem and I. K. Park, “Robust light field depth estimation for noisy scene with occlusion”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4396–4404. DOI: https://doi.org/10.1109/CVPR.2016.476.

[129] N. Zeller, F. Quint, and U. Stilla, “Depth estimation and camera calibration of a focused plenoptic camera for visual odometry”, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 118, pp. 83–100, 2016. DOI: https://doi.org/10.1016/j.isprsjprs.2016.04.010.

[130] Y. Zhang, H. Lv, Y. Liu, *et al.*, “Light-field depth estimation via epipolar plane image analysis and locally linear embedding”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 739–747, 2016. DOI: https://doi.org/10.1109/TCSVT.2016.2555778.

[131] W. Zhou, A. Lumsdaine, and L. Lin, “Depth estimation with cascade occlusion culling filter for light-field cameras”, in *2016 23rd International Conference on Pattern Recognition (ICPR)*, IEEE, 2016, pp. 1887–1892. DOI: https://doi.org/10.1109/ICPR.2016.7899912.

[132] D. Antensteiner, S. Štolc, and R. Huber-Mörk, “Depth estimation with light field and photometric stereo data using energy minimization”, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 21st Iberoamerican Congress, CIARP 2016, Lima, Peru, November 8–11, 2016, Proceedings 21*, Springer International Publishing, 2017, pp. 175–183. DOI: https://doi.org/10.1007/978-3-319-52277-7_22.

[133] J. Fan and Y.-H. Yang, “Depth estimation of semi-submerged objects using a light-field camera”, in *2017 14th Conference on Computer and Robot Vision (CRV)*, IEEE, 2017, pp. 80–86. DOI: https://doi.org/10.1109/CRV.2017.44.

[134] S. Heber, W. Yu, and T. Pock, “Neural epi-volume networks for shape from light field”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2252–2260. DOI: https://doi.org/10.1109/ICCV.2017.247.

[135] Q. Hou and C. Jung, “Occlusion robust light field depth estimation using segmentation guided bilateral filtering”, in *2017 IEEE International Symposium on Multimedia (ISM)*, IEEE, 2017, pp. 14–18. DOI: https://doi.org/10.1109/ISM.2017.13.

[136] C.-T. Huang, “Robust pseudo random fields for light-field stereo matching”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 11–19. DOI: https://doi.org/10.1109/ICCV.2017.11.

[137] O. Johannsen, K. Honauer, B. Goldluecke, *et al.*, “A taxonomy and evaluation of dense light field depth estimation algorithms”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 82–99. DOI: https://doi.org/10.1109/CVPRW.2017.226.

[138] J. Y. Lee and R.-H. Park, “Depth estimation from light field by accumulating binary maps based on foreground–background separation”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 955–964, 2017. DOI: https://doi.org/10.1109/JSTSP.2017.2747154.

[139] W.-Y. Lee, C.-Y. Li, and J.-Y. Yen, “Integrating wavelet transformation with markov random field analysis for the depth estimation of light-field images”, *IET Computer Vision*, vol. 11, no. 5, pp. 358–367, 2017. DOI: https://doi.org/10.1049/iet-cvi.2016.0151.

[140] F. Liu, G. Hou, Z. Sun, and T. Tan, “High quality depth map estimation of object surface from light-field images”, *Neurocomputing*, vol. 252, pp. 3–16, 2017. DOI: https://doi.org/10.1016/j.neucom.2016.09.136.

[141] Y. Luo, W. Zhou, J. Fang, L. Liang, H. Zhang, and G. Dai, “Epi-patch based convolutional neural network for depth estimation on 4d light field”, in *International Conference on Neural Information Processing*, Springer, 2017, pp. 642–652. DOI: https://doi.org/10.1007/978-3-319-70090-8_65.

[142] Z. Ma, Z. Cen, and X. Li, “Depth estimation algorithm for light field data by epipolar image analysis and region interpolation”, *Applied Optics*, vol. 56, no. 23, pp. 6603–6610, 2017. DOI: https://doi.org/10.1364/AO.56.006603.

[143] J. Navarro and A. Buades, “Robust and dense depth estimation for light field images”, *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1873–1886, 2017. DOI: https://doi.org/10.1109/TIP.2017.2666041.

[144] I. K. Park, K. M. Lee, *et al.*, “Robust light field depth estimation using occlusion-noise aware data costs”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2484–2497, 2017. DOI: https://doi.org/10.1109/TPAMI.2017.2746858.

[145] Y. Qin, X. Jin, Y. Chen, and Q. Dai, “Enhanced depth estimation for handheld light field cameras”, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 2032–2036. DOI: https://doi.org/10.1109/ICASSP.2017.7952513.

[146] Y. Qin, X. Jin, and Q. Dai, "Gpu-based depth estimation for light field images", in *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, IEEE, 2017, pp. 640–645. DOI: https://doi.org/10.1109/ISPACS.2017.8266556.

[147] H. Sheng, S. Zhang, X. Cao, Y. Fang, and Z. Xiong, "Geometric occlusion analysis in depth estimation using integral guided filter for light-field image", *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5758–5771, 2017. DOI: https://doi.org/10.1109/TIP.2017.2745100.

[148] L. Si and Q. Wang, "Dense depth-map estimation and geometry inference from light fields via global optimization", in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, Springer, 2017, pp. 83–98. DOI: https://doi.org/10.1007/978-3-319-54187-7_6.

[149] J. Tian, Z. Murez, T. Cui, Z. Zhang, D. Kriegman, and R. Ramamoorthi, "Depth and image restoration from light field in a scattering medium", in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2401–2410. DOI: https://doi.org/10.1109/ICCV.2017.263.

[150] T. Tomioka, K. Mishiba, Y. Oyamada, and K. Kondo, "Depth map estimation using census transform for light field cameras", *IEICE TRANSACTIONS on Information and Systems*, vol. 100, no. 11, pp. 2711–2720, 2017. DOI: https://doi.org/10.1587/transinf.2017EDP7052.

[151] R. Verhack, T. Sikora, L. Lange, R. Jongebloed, G. Van Wallendael, and P. Lambert, "Steered mixture-of-experts for light field coding, depth estimation, and processing", in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 1183–1188. DOI: https://doi.org/10.1109/ICME.2017.8019442.

[152] H. Zhu, Q. Wang, and J. Yu, "Occlusion-model guided antiocclusion depth estimation in light field", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 965–978, 2017. DOI: https://doi.org/10.1109/JSTSP.2017.2730818.

[153] M. Z. Alam and B. K. Gunturk, "Hybrid light field imaging for improved spatial resolution and depth range", *Machine Vision and Applications*, vol. 29, no. 1, pp. 11–22, 2018. DOI: https://doi.org/10.1007/s00138-017-0862-2.

[154] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9145–9154. DOI: https://doi.org/10.1109/CVPR.2018.00953.

[155] Z. Cai, X. Liu, X. Peng, and B. Z. Gao, "Ray calibration and phase mapping for structured-light-field 3d reconstruction", *Optics Express*, vol. 26, no. 6, pp. 7598–7613, 2018. DOI: https://doi.org/10.1364/OE.26.007598.

[156] Z. Cai, X. Liu, X. Peng, and B. Z. Gao, "Universal phase-depth mapping in a structured light field", *Applied Optics*, vol. 57, no. 1, A26–A32, 2018. DOI: https://doi.org/10.1364/AO.57.000A26.

[157] W. Chantara and Y.-S. Ho, "Initial depth estimation using adaptive window size with light field image", in *2018 International Workshop on Advanced Image Technology (IWAIT)*, IEEE, 2018, pp. 1–3. DOI: https://doi.org/10.1109/IWAIT.2018.8369722.

[158] L.-D. Chen, Y.-T. Lu, Y.-L. Hiao, B.-H. Yang, W.-C. Chen, and C.-T. Huang, "A 95pj/label wide-range depth-estimation processor for full-hd light-field applications on fpga", in *2018 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, IEEE, 2018, pp. 261–262. DOI: https://doi.org/10.1109/ASSCC.2018.8579289.

[159] J. Chen, J. Hou, Y. Ni, and L.-P. Chau, "Accurate light field depth estimation with superpixel regularization over partially occluded regions", *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4889–4900, 2018. DOI: https://doi.org/10.1109/TIP.2018.2839524.

[160] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011. DOI: https://doi.org/10.1109/CVPR.2018.00214.

[161] Q. Han and C. Jung, "Guided filtering based data fusion for light field depth estimation with l0 gradient minimization", *Journal of Visual Communication and Image Representation*, vol. 55, pp. 449–456, 2018. DOI: https://doi.org/10.1016/j.jvcir.2018.06.020.

[162] X. Huang, P. An, L. Shen, and R. Ma, "Efficient light field images compression method based on depth estimation and optimization", *IEEE Access*, vol. 6, pp. 48 984–48 993, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2867862.

[163] X. Huang, P. An, L. Shan, R. Ma, and L. Shen, "View synthesis for light field coding using depth estimation", in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2018, pp. 1–6. DOI: https://doi.org/10.1109/ICME.2018.8486515.

[164] A. Ivan, I. Kyu Park, *et al.*, "Light field depth estimation on off-the-shelf mobile gpu", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 634–643. DOI: https://doi.org/10.1109/CVPRW.2018.00106.

[165] H.-G. Jeon, J. Park, G. Choe, *et al.*, "Depth from a light field image with learning-based matching costs", *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 297–310, 2018. DOI: https://doi.org/10.1109/TPAMI.2018.2794979.

[166] X. Jiang, M. Le Pendu, and C. Guillemot, "Depth estimation with occlusion handling from a sparse set of light field views", in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 634–638. DOI: https://doi.org/10.1109/ICIP.2018.8451466.

[167] M. Ko, D. Kim, M. Kim, and K. Kim, "Illumination-insensitive skin depth estimation from a light-field camera based on cgans toward haptic palpation", *Electronics*, vol. 7, no. 11, p. 336, 2018. DOI: https://doi.org/10.3390/electronics7110336.

[168] D. Lee, H. Park, I. K. Park, and K. M. Lee, “Joint blind motion deblurring and depth estimation of light field”, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 288–303. DOI: `https://doi.org/10.1007/978-3-030-01270-0_18`.

[169] J. Y. Lee and R.-H. Park, “Reduction of aliasing artifacts by sign function approximation in light field depth estimation based on foreground–background separation”, *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1750–1754, 2018. DOI: `https://doi.org/10.1109/LSP.2018.2874304`.

[170] X. Liu, Q. Wang, Z. Ma, *et al.*, “Depth estimation and multi-view spectral images based on compressive sensing light field reconstruction”, in *3D Image Acquisition and Display: Technology, Perception and Applications*, Optical Society of America, 2018, 3Tu3E–5. DOI: `https://doi.org/10.1364/3D.2018.3Tu3E.5`.

[171] H. Lu, Y. Li, H. Kim, and S. Serikawa, “Underwater light field depth map restoration using deep convolutional neural fields”, in *Artificial intelligence and robotics*, Springer, 2018, pp. 305–312. DOI: `https://doi.org/10.1007/978-3-319-69877-9_33`.

[172] H. Ma, H. Li, Z. Qian, S. Shi, and T. Mu, “Vommanet: An end-to-end network for disparity estimation from reflective and texture-less light field images”, *arXiv preprint arXiv:1811.07124*, 2018. DOI: `https://doi.org/10.48550/arXiv.1811.07124`.

[173] J.-H. Mun and Y.-S. Ho, “Depth estimation from light field images via convolutional residual network”, in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2018, pp. 1495–1498. DOI: `https://doi.org/10.23919/APSIPA.2018.8659639`.

[174] I. K. Park *et al.*, “Cost aggregation benchmark for light field depth estimation”, *Journal of Visual Communication and Image Representation*, vol. 56, pp. 38–51, 2018. DOI: `https://doi.org/10.1016/j.jvcir.2018.08.015`.

[175] J. Peng, Z. Xiong, D. Liu, and X. Chen, “Unsupervised depth estimation from light field using a convolutional neural network”, in *2018 International Conference on 3D Vision (3DV)*, IEEE, 2018, pp. 295–303. DOI: `https://doi.org/10.1109/3DV.2018.00042`.

[176] S. Pertuz, E. Pulido-Herrera, and J.-K. Kamarainen, “Focus model for metric depth estimation in standard plenoptic cameras”, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 144, pp. 38–47, 2018. DOI: `https://doi.org/10.1016/j.isprsjprs.2018.06.020`.

[177] H. Sheng, P. Zhao, S. Zhang, J. Zhang, and D. Yang, “Occlusion-aware depth estimation for light field using multi-orientation epis”, *Pattern Recognition*, vol. 74, pp. 587–599, 2018. DOI: `https://doi.org/10.1016/j.patcog.2017.09.010`.

[178] W. Zhou, L. Liang, H. Zhang, A. Lumsdaine, and L. Lin, “Scale and orientation aware epi-patch learning for light field depth estimation”, in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 2018, pp. 2362–2367. DOI: `https://doi.org/10.1109/ICPR.2018.8545490`.

[179] W. Ai, S. Xiang, and L. Yu, “Robust depth estimation for multi-occlusion in light-field images”, *Optics Express*, vol. 27, no. 17, pp. 24 793–24 807, 2019. DOI: `https://doi.org/10.1364/OE.27.024793`.

[180] Y. Anisimov, O. Wasenmüller, and D. Stricker, “A compact light field camera for real-time depth estimation”, in *International Conference on Computer Analysis of Images and Patterns*, Springer, 2019, pp. 52–63. DOI: `https://doi.org/10.1007/978-3-030-29888-3_5`.

[181] Y. Anisimov, O. Wasenmüller, and D. Stricker, “Rapid light field depth estimation with semi-global matching”, in *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, IEEE, 2019, pp. 109–116. DOI: `https://doi.org/10.1109/ICCP48234.2019.8959680`.

[182] Z. Cai, X. Liu, G. Pedrini, W. Osten, and X. Peng, “Accurate depth estimation in structured light fields”, *Optics Express*, vol. 27, no. 9, pp. 13 532–13 546, 2019. DOI: `https://doi.org/10.1364/OE.27.013532`.

[183] A. Chuchvara, A. Barsi, and A. Gotchev, “Fast and accurate depth estimation from sparse light fields”, *IEEE Transactions on Image Processing*, vol. 29, pp. 2492–2506, 2019. DOI: `https://doi.org/10.1109/TIP.2019.2959233`.

[184] D. G. Dansereau, B. Girod, and G. Wetzstein, “Liff: Light field features in scale and depth”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8042–8051. DOI: `https://doi.org/10.1109/CVPR.2019.00823`.

[185] C. Domínguez Conde, J. P. Lüke, and F. Rosa González, “Implementation of a depth from light field algorithm on fpga”, *Sensors*, vol. 19, no. 16, p. 3562, 2019. DOI: `https://doi.org/10.3390/s19163562`.

[186] Á. Faluvégi, Q. Bolseé, S. Nedevschi, V.-T. Dădârlat, and A. Munteanu, “A 3d convolutional neural network for light field depth estimation”, in *2019 International Conference on 3D Immersion (IC3D)*, IEEE, 2019, pp. 1–5. DOI: `https://doi.org/10.1109/IC3D48390.2019.8975996`.

[187] M. Ghorai and A. Munteanu, “Depth estimation with occlusion prediction in light field images”, in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 1049–1053. DOI: `https://doi.org/10.1109/ICIP.2019.8803722`.

[188] Z. Guo, J. Wu, X. Chen, *et al.*, “Accurate light field depth estimation using multi-orientation partial angular coherence”, *IEEE Access*, vol. 7, pp. 169 123–169 132, 2019. DOI: `https://doi.org/10.1109/ACCESS.2019.2954892`.

[189] X. Huang, P. An, F. Cao, D. Liu, and Q. Wu, “Light-field compression using a pair of steps and depth estimation”, *Optics express*, vol. 27, no. 3, pp. 3557–3573, 2019. DOI: `https://doi.org/10.1364/OE.27.003557`.

[190] X. Jiang, J. Shi, and C. Guillemot, “A learning based depth estimation framework for 4d densely and sparsely sampled light fields”, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2257–2261. DOI: `https://doi.org/10.1109/ICASSP.2019.8683773`.

[191] M. Ko, D. Kim, and K. Kim, "Accurate depth estimation of skin surface using a light-field camera toward dynamic haptic palpation", *Skin Research and Technology*, vol. 25, no. 4, pp. 469–481, 2019. DOI: https://doi.org/10.1111/srt.12675.

[192] T. Leistner, H. Schilling, R. Mackowiak, S. Gumhold, and C. Rother, "Learning to think outside the box: Wide-baseline light field depth estimation with epi-shift", in *2019 International Conference on 3D Vision (3DV)*, IEEE, 2019, pp. 249–257. DOI: https://doi.org/10.1109/3DV.2019.00036.

[193] J. Li and X. Jin, "Frequency descriptor based light field depth estimation", in *2019 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2019, pp. 1–4. DOI: https://doi.org/10.1109/VCIP47243.2019.8965944.

[194] F. Liu, S. Zhou, Y. Wang, G. Hou, Z. Sun, and T. Tan, "Binocular light-field: Imaging theory and occlusion-robust depth perception application", *IEEE Transactions on Image Processing*, vol. 29, pp. 1628–1640, 2019. DOI: https://doi.org/10.1109/TIP.2019.2943019.

[195] R. Lourenco, P. Assuncao, L. Tavera, L. A. Thomaz, R. Pinto, and S. Faria, "Edge reconstruction method to improve depth estimation from light fields", in *EURASIP European Light Field Imaging Workshop ELFI*, 2019.

[196] Y. Mo, J. Yang, C. Xiao, and W. An, "Toward real-world light field depth estimation: A noise-aware paradigm using multi-stereo disparity integration", *IEEE Access*, vol. 7, pp. 94 391–94 399, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2928006.

[197] L. Palmieri, G. Scrofani, N. Incardona, G. Saavedra, M. Martínez-Corral, and R. Koch, "Robust depth estimation for light field microscopy", *Sensors*, vol. 19, no. 3, p. 500, 2019. DOI: https://doi.org/10.3390/s19030500.

[198] X. Pan, T. Zhang, and H. Wang, "A method for handling multi-occlusion in depth estimation of light field", in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 1069–1073. DOI: https://doi.org/10.1109/ICIP.2019.8804273.

[199] I. Schiopu and A. Munteanu, "Deep-learning-based depth estimation from light field images", *Electronics Letters*, vol. 55, no. 20, pp. 1086–1088, 2019. DOI: https://doi.org/10.1049/el.2019.2073.

[200] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views", *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5867–5880, 2019. DOI: https://doi.org/10.1109/TIP.2019.2923323.

[201] A. Stacey, W. Maddern, and S. Singh, "Fast light field disparity estimation via a parallel filtered cost volume approach", in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, Springer, 2019, pp. 256–268. DOI: https://doi.org/10.1007/978-3-030-20890-5_17.

[202] X. Wang, W. Liu, Y. Sun, L. Yang, Z. Qin, and Z. Zheng, "A learning-based method using epipolar geometry for light field depth estimation", in *Optoelectronic Imaging and Multimedia Technology VI*, SPIE, vol. 11187, 2019, pp. 24–32. DOI: https://doi.org/10.1117/12.2537208.

[203] T. Yan, F. Zhang, Y. Mao, H. Yu, X. Qian, and R. W. Lau, "Depth estimation from a light field image pair with a generative model", *IEEE Access*, vol. 7, pp. 12 768–12 778, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2893354.

[204] C. Yang, Z. Liu, K. Di, Y. Wang, and M. Peng, "Improved camera distortion correction and depth estimation for lenslet light field camera", *Photogrammetric Engineering & Remote Sensing*, vol. 85, no. 3, pp. 197–208, 2019. DOI: https://doi.org/10.14358/PERS.85.3.197.

[205] Y. Zhang, W. Dai, M. Xu, J. Zou, X. Zhang, and H. Xiong, "Depth estimation from light field using graph-based structure-aware analysis", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4269–4283, 2019. DOI: https://doi.org/10.1109/TCSVT.2019.2954948.

[206] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image", *IEEE Transactions on Image Processing*, vol. 29, pp. 1606–1617, 2019. DOI: https://doi.org/10.1109/TIP.2019.2944343.

[207] Z. Cai, X. Liu, G. Pedrini, W. Osten, and X. Peng, "Light-field depth estimation considering plenoptic imaging distortion", *Optics Express*, vol. 28, no. 3, pp. 4156–4168, 2020. DOI: https://doi.org/10.1364/OE.385285.

[208] C. Guo, J. Jin, J. Hou, and J. Chen, "Accurate light field depth estimation via an occlusion-aware network", in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2020, pp. 1–6. DOI: https://doi.org/10.1109/ICME46284.2020.9102829.

[209] Z. Hu, Y. Y. Chung, W. Ouyang, X. Chen, and Z. Chen, "Light field reconstruction using hierarchical features fusion", *Expert Systems with Applications*, vol. 151, p. 113 394, 2020. DOI: https://doi.org/10.1016/j.eswa.2020.113394.

[210] Z. Huang, J. A. Fessler, T. B. Norris, and I. Y. Chun, "Light-field reconstruction and depth estimation from focal stack images using convolutional neural networks", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 8648–8652. DOI: https://doi.org/10.1109/ICASSP40776.2020.9053586.

[211] D. Jin, A. Zhang, J. Wu, G. Wu, H. Wang, and L. Fang, "All-in-depth via cross-baseline light field camera", in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3559–3567. DOI: https://doi.org/10.1145/3394171.3413974.

[212] N. Khan, M. H. Kim, and J. Tompkin, "View-consistent 4d light field depth estimation", *arXiv preprint arXiv:2009.04065*, 2020. DOI: https://doi.org/10.48550/arXiv.2009.04065.

[213] J. Li and X. Jin, "Epi-neighborhood distribution based light field depth estimation", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 2003–2007. DOI: https://doi.org/10.1109/ICASSP40776.2020.9053664.

[214] Y. Li, L. Zhang, Q. Wang, and G. Lafruit, “Manet: Multi-scale aggregated network for light field depth estimation”, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 1998–2002. DOI: `https://doi.org/10.1109/ICASSP40776.2020.9053532`.

[215] X. Liu, D. Fu, C. Wu, and Z. Si, “The depth estimation method based on double-cues fusion for light field images”, in *Proceedings of the 11th International Conference on Modelling, Identification and Control (ICMIC2019)*, Springer, 2020, pp. 719–726. DOI: `https://doi.org/10.1007/978-981-15-0474-7_67`.

[216] D. Liu, Y. Huang, Q. Wu, R. Ma, and P. An, “Multi-angular epipolar geometry based light field angular reconstruction network”, *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1507–1522, 2020. DOI: `https://doi.org/10.1109/TCI.2020.3037413`.

[217] K. Mishiba, “Fast depth estimation for light field cameras”, *IEEE Transactions on Image Processing*, vol. 29, pp. 4232–4242, 2020. DOI: `https://doi.org/10.1109/TIP.2020.2970814`.

[218] J. Peng, Z. Xiong, Y. Wang, Y. Zhang, and D. Liu, “Zero-shot depth estimation from light field using a convolutional neural network”, *IEEE Transactions on Computational Imaging*, vol. 6, pp. 682–696, 2020. DOI: `https://doi.org/10.1109/TCI.2020.2967148`.

[219] S. Rogge, I. Schiopu, and A. Munteanu, “Depth estimation for light-field images using stereo matching and convolutional neural networks”, *Sensors*, vol. 20, no. 21, p. 6188, 2020. DOI: `https://doi.org/10.3390/s20216188`.

[220] W. Si, “An effective occlusion edge prediction method in light field depth estimation”, in *Frontiers in Cyber Security: Third International Conference, FCS 2020, Tianjin, China, November 15–17, 2020, Proceedings*, Springer Nature, vol. 1286, 2020, p. 403. DOI: `https://doi.org/10.1007/978-981-15-9739-8_31`.

[221] T.-H. Tran, G. Mammadov, and S. Simon, “Gvld: A fast and accurate gpu-based variational light-field disparity estimation approach”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2562–2574, 2020. DOI: `https://doi.org/10.1109/TCSVT.2020.3028258`.

[222] V. Van Duong, T. N. Huu, and B. Jeon, “Robust light field depth estimation with occlusion based on spatial and spectral entropies data costs”, in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 2631–2635. DOI: `https://doi.org/10.1109/ICIP40778.2020.9191162`.

[223] X. Wang, C. Tao, R. Wu, *et al.*, “Light-field-depth-estimation network based on epipolar geometry and image segmentation”, *JOSA A*, vol. 37, no. 7, pp. 1236–1243, 2020. DOI: `https://doi.org/10.1364/JOSAA.388555`.

[224] C. Yang, Z. Liu, K. Di, C. Hu, Y. Wang, and W. Liang, “Improved depth estimation for occlusion scenes using a light-field camera”, *Photogrammetric Engineering & Remote Sensing*, vol. 86, no. 7, pp. 443–456, 2020. DOI: `https://doi.org/10.14358/PERS.86.7.443`.

[225] J.-S. Yun and J.-Y. Sim, “Deep learning based depth estimation and reconstruction of light field images”, in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 1252–1256.

[226] C. Zhou, Q. Zhang, B. Wang, Y. Du, T. Yan, and W. Si, “An effective occlusion edge prediction method in light field depth estimation”, in *International Conference on Frontiers in Cyber Security*, Springer, 2020, pp. 403–416. DOI: `https://doi.org/10.1007/978-981-15-9739-8_31`.

[227] S. T. Digumarti, J. Daniel, A. Ravendran, R. Griffiths, and D. G. Dansereau, “Unsupervised learning of depth estimation and visual odometry for sparse light field cameras”, in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 278–285. DOI: `https://doi.org/10.1109/IROS51168.2021.9636570`.

[228] Y. Du, Q. Zhang, D. Hua, *et al.*, “Eanet: Depth estimation based on epi of light field”, *BioMed Research International*, vol. 2021, 2021. DOI: `https://doi.org/10.1155/2021/8293151`.

[229] L. Han, X. Huang, Z. Shi, and S. Zheng, “Depth estimation from light field geometry using convolutional neural networks”, *Sensors*, vol. 21, no. 18, p. 6061, 2021. DOI: `https://doi.org/10.3390/s21186061`.

[230] L. Han, X. Huang, Z. Shi, and S. Zheng, “Learning depth from light field via deep convolutional neural network”, in *Big Data and Security: Second International Conference, ICBDS 2020, Singapore, Singapore, December 20–22, 2020, Revised Selected Papers 2*, Springer, 2021, pp. 485–496. DOI: `https://doi.org/10.1007/978-981-16-3150-4_40`.

[231] K. Han, W. Xiang, E. Wang, and T. Huang, “A novel occlusion-aware vote cost for light field depth estimation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. DOI: `https://doi.org/10.1109/TPAMI.2021.3105523`.

[232] Z. Huang, X. Hu, Z. Xue, W. Xu, and T. Yue, “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. DOI: `https://doi.org/10.1109/ICCV48922.2021.00626`.

[233] N. Khan, M. H. Kim, and J. Tompkin, “Edge-aware bidirectional diffusion for dense depth estimation from light fields”, *arXiv preprint arXiv:2107.02967*, 2021. DOI: `https://doi.org/10.48550/arXiv.2107.02967`.

[234] T. Kinoshita and S. Ono, “Depth estimation from 4d light field videos”, in *International Workshop on Advanced Imaging Technology (IWAIT) 2021*, SPIE, vol. 11766, 2021, pp. 56–61. DOI: `https://doi.org/10.1117/12.2591012`.

[235] J. Y. Lee, R.-H. Park, and J. Kim, “Occlusion handling by successively excluding foregrounds for light field depth estimation based on foreground-background separation”, *IEEE Access*, vol. 9, pp. 103 927–103 936, 2021. DOI: `https://doi.org/10.1109/ACCESS.2021.3098819`.

[236] K. Li, J. Zhang, J. Gao, and M. Qi, “Self-supervised light field depth estimation using epipolar plane images”, in *2021 International Conference on 3D Vision (3DV)*, IEEE, 2021, pp. 731–740. DOI: `https://doi.org/10.1109/3DV53792.2021.00082`.

[237] D. Ma and A. Lumsdaine, “Fast and efficient neural network for light field disparity estimation”, in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 2920–2926. DOI: `https://doi.org/10.1109/ICPR48806.2021.9412856`.

[238] S. Ma, Z. Guo, J. Wu, *et al.*, “Occlusion-aware light field depth estimation using side window angular coherence”, *Applied Optics*, vol. 60, no. 2, pp. 392–404, 2021. DOI: `https://doi.org/10.1364/AO.411070`.

[239] Y. Mo, C. Xiao, J. Yang, and W. An, “Baseline-adaptive light field depth estimation based on stereo matching”, in *2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*, IEEE, 2021, pp. 841–845. DOI: `https://doi.org/10.1109/ISCIPT53667.2021.00176`.

[240] S. Nehra, T. Das, S. Chakraborty, P. Biswas, and J. Mukhopadhyay, “Disparity based depth estimation using light field camera”, in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, 2021, pp. 1–9. DOI: `https://doi.org/10.1145/3490035.3490297`.

[241] Y. Piao, Y. Zhang, M. Zhang, and X. Ji, “Dynamic fusion network for light field depth estimation”, *arXiv preprint arXiv:2104.05969*, 2021. DOI: `https://doi.org/10.1007/978-3-030-88007-1_1`.

[242] Y. Piao, X. Ji, M. Zhang, and Y. Zhang, “Learning multi-modal information for robust light field depth estimation”, *arXiv preprint arXiv:2104.05971*, 2021. DOI: `https://doi.org/10.48550/arXiv.2104.05971`.

[243] T. Sasaki, C. Hashemi, and J. R. Leger, “Depth estimation of non-line-of-sight objects from scattered light enhanced by the infrared light field”, in *Computational Optical Sensing and Imaging*, Optical Society of America, 2021, CW5B–5. DOI: `https://doi.org/10.1364/COSI.2021.CW5B.5`.

[244] W. Wang, Y. Lin, and S. Zhang, “Enhanced spinning parallelogram operator combining color constraint and histogram integration for robust light field depth estimation”, *IEEE Signal Processing Letters*, vol. 28, pp. 1080–1084, 2021. DOI: `https://doi.org/10.1109/LSP.2021.3079844`.

[245] S. Xiang, L. Liu, H. Deng, J. Wu, Y. Yang, and L. Yu, “Fast depth estimation with cost minimization for structured light field”, *Optics Express*, vol. 29, no. 19, pp. 30 077–30 093, 2021. DOI: `https://doi.org/10.1364/OE.434548`.

[246] Y. Zhang, Y. Piao, X. Ji, and M. Zhang, “Dynamic fusion network for light field depth estimation”, in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Springer, 2021, pp. 3–15. DOI: `https://doi.org/10.1007/978-3-030-88007-1_1`.

[247] L. Zhang, H. Deng, S. Xiang, and S. Li, “Light field depth estimation based on occlusion optimization”, in *2021 33rd Chinese Control and Decision Conference (CCDC)*, IEEE, 2021, pp. 1635–1639. DOI: `https://doi.org/10.1109/CCDC52312.2021.9602123`.

[248] Y. Anisimov, J. R. Rambach, and D. Stricker, “Nonlinear optimization of light field point cloud”, *Sensors*, vol. 22, no. 3, p. 814, 2022. DOI: https://doi.org/10.3390/s22030814.

[249] W. Chao, X. Wang, Y. Wang, L. Chang, and F. Duan, “Learning subpixel disparity distribution for light field depth estimation”, *arXiv preprint arXiv:2208.09688*, 2022. DOI: https://doi.org/10.1109/TCI.2023.3336184.

[250] M. Gao, H. Deng, S. Xiang, J. Wu, and Z. He, “Epi light field depth estimation based on a directional relationship model and multiviewpoint attention mechanism”, *Sensors*, vol. 22, no. 16, p. 6291, 2022. DOI: https://doi.org/10.3390/s22166291.

[251] L. Han, Z. Shi, S. Zheng, X. Huang, and M. Xu, “Light-field depth estimation using rnn and crf”, in *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, IEEE, 2022, pp. 725–729. DOI: https://doi.org/10.1109/ICIVC55077.2022.9886991.

[252] A. Hassan, M. Sjöström, T. Zhang, and K. Egiazarian, “Light-weight epinet architecture for fast light field disparity estimation”, in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2022, pp. 1–5. DOI: https://doi.org/10.1109/MMSP55362.2022.9949378.

[253] T. Iwatsuki, K. Takahashi, and T. Fujii, “Unsupervised disparity estimation from light field using plug-and-play weighted warping loss”, *Signal Processing: Image Communication*, vol. 107, p. 116 764, 2022. DOI: https://doi.org/10.1016/j.image.2022.116764.

[254] J. Jin and J. Hou, “Occlusion-aware unsupervised learning of depth from 4-d light fields”, *IEEE Transactions on Image Processing*, vol. 31, pp. 2216–2228, 2022. DOI: https://doi.org/10.1109/TIP.2022.3154288.

[255] T. Leistner, R. Mackowiak, L. Ardizzone, U. Köthe, and C. Rother, “Towards multimodal depth estimation from light fields”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 953–12 961. DOI: https://doi.org/10.1109/CVPR52688.2022.01261.

[256] L. Lin, Q. Li, B. Gao, Y. Yan, W. Zhou, and E. E. Kuruoglu, “Unsupervised learning of light field depth estimation with spatial and angular consistencies”, *Neurocomputing*, vol. 501, pp. 113–122, 2022. DOI: https://doi.org/10.1016/j.neucom.2022.06.011.

[257] R. Lourenco, L. A. Thomaz, E. A. da Silva, and S. M. de Faria, “Enhanced local optimization framework for light field disparity estimation”, in *2022 10th European Workshop on Visual Information Processing (EUVIP)*, IEEE, 2022, pp. 1–6. DOI: https://doi.org/10.1109/EUVIP53989.2022.9922672.

[258] R. M. Lourenco, L. M. Tavora, P. A. Assuncao, L. A. Thomaz, R. Fonseca-Pinto, and S. M. Faria, “Enhancement of light field disparity maps by reducing the silhouette effect and plane noise”, *Multidimensional Systems and Signal Processing*, vol. 33, no. 2, pp. 1–33, 2022. DOI: https://doi.org/10.1007/s11045-021-00807-7.

[259] E. Martínez, E. Vargas, and H. Arguello, "Fast disparity estimation from a single compressed light field measurement", in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 1991–1995. DOI: https://doi.org/10.23919/EUSIPCO55093.2022.9909979.

[260] Y. Shen, Y. Liu, Y. Tian, Z. Liu, and F. Wang, "A new parallel intelligence based light field dataset for depth refinement and scene flow estimation", *Sensors*, vol. 22, no. 23, p. 9483, 2022. DOI: https://doi.org/10.3390/s22239483.

[261] V. Van Duong, T. N. Huu, J. Yim, and B. Jeon, "Lfdenet: Light field depth estimation network based on hybrid data representation", in *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, IEEE, 2022, pp. 1–4. DOI: https://doi.org/10.1109/BMSB55706.2022.9828626.

[262] Y. Wang, L. Wang, G. Wu, *et al.*, "Disentangling light fields for super-resolution and disparity estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 425–443, 2022. DOI: https://doi.org/10.1109/TPAMI.2022.3152488.

[263] Y. Wang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, "Occlusion-aware cost constructor for light field depth estimation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 809–19 818. DOI: https://doi.org/10.1109/CVPR52688.2022.01919.

[264] W. Yan, X. Zhang, H. Chen, C. Ling, and D. Wang, "Light field depth estimation based on channel attention and edge guidance", in *2022 China Automation Congress (CAC)*, IEEE, 2022, pp. 2595–2600. DOI: https://doi.org/10.1109/CAC57257.2022.10054964.

[265] Y. Zhao, Z. Cui, R. Chen, D. Yang, and H. Sheng, "Lf-dwnet: Robust depth estimation network for light field with disparity warping", in *International Conference on Wireless Algorithms, Systems, and Applications*, Springer, 2022, pp. 291–302. DOI: https://doi.org/10.1007/978-3-031-19214-2_24.

[266] Z. Zhao, S. Cheng, and L. Li, "Robust depth estimation on real-world light field images using gaussian belief propagation", *Image and Vision Computing*, vol. 122, p. 104 447, 2022. DOI: https://doi.org/10.1016/j.imavis.2022.104447.

[267] P. Zhou, X. Liu, J. Jin, Y. Zhang, and J. Hou, "Light field depth estimation based on stitched-epi", *arXiv preprint arXiv:2203.15201*, 2022. DOI: https://doi.org/10.48550/arXiv.2203.15201.

[268] Y. Zihang, Z. Yueming, and D. Huiping, "Multi-cue depth estimation from epipolar plane image", in *2022 34th Chinese Control and Decision Conference (CCDC)*, IEEE, 2022, pp. 2619–2625. DOI: https://doi.org/10.1109/CCDC55256.2022.10034099.

[269] W. Chao, X. Wang, Y. Kan, and F. Duan, “Contextnet: Learning context information for texture-less light field depth estimation”, in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Springer, 2023, pp. 15–27. DOI: https://doi.org/10.1007/978-981-99-8537-1_2.

[270] W. Chao, X. Wang, Y. Wang, G. Wang, and F. Duan, “Learning sub-pixel disparity distribution for light field depth estimation”, *IEEE Transactions on Computational Imaging*, vol. 9, pp. 1126–1138, 2023. DOI: https://doi.org/10.1109/TCI.2023.3336184.

[271] W. Chao, F. Duan, X. Wang, Y. Wang, and G. Wang, “Occcasnet: Occlusion-aware cascade cost volume for light field depth estimation”, *arXiv preprint arXiv:2305.17710*, 2023. DOI: https://doi.org/10.48550/arXiv.2305.17710.

[272] R. Chen, H. Sheng, D. Yang, S. Wang, Z. Cui, and R. Cong, “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*, vol. 37, 2023, pp. 331–339. DOI: https://doi.org/10.1609/aaai.v37i1.25106.

[273] Z. Cui, H. Sheng, D. Yang, S. Wang, R. Chen, and W. Ke, “Light field depth estimation for non-lambertian objects via adaptive cross operator”, *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. DOI: https://doi.org/10.1109/TCSVT.2023.3292884.

[274] C. Fu, H. Yuan, H. Xu, H. Zhang, and L. Shen, “Tmso-net: Texture adaptive multi-scale observation for light field image depth estimation”, *Journal of Visual Communication and Image Representation*, vol. 90, p. 103 731, 2023. DOI: https://doi.org/10.1016/j.jvcir.2022.103731.

[275] L. Han, S. Zheng, Z. Shi, and M. Xia, “Exploiting sequence analysis for accurate light-field depth estimation”, *IEEE Access*, 2023. DOI: https://doi.org/10.1109/ACCESS.2023.3296800.

[276] D. Hua, Q. Zhang, W. Liao, B. Wang, and T. Yan, “Cattnet: A compound attention network for depth estimation of light field images”, *J. Inf. Process. Syst*, vol. 19, pp. 483–497, 2023. DOI: https://doi.org/10.3745/JIPS.02.0201.

[277] P. Li, J. Zhao, J. Wu, *et al.*, “Opal: Occlusion pattern aware loss for unsupervised light field disparity estimation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. DOI: https://doi.org/10.1109/TPAMI.2023.3296600.

[278] W. Liao, X. Bai, Q. Zhang, *et al.*, “Decoupled and reparameterized compound attention-based light field depth estimation network”, *IEEE Access*, vol. 11, pp. 130 119–130 130, 2023. DOI: https://doi.org/10.1109/ACCESS.2023.3334640.

[279] Y. Liu, M. Aleksandrov, Z. Hu, *et al.*, “Accurate light field depth estimation under occlusion”, *Pattern Recognition*, vol. 138, p. 109 415, 2023. DOI: https://doi.org/10.1016/j.patcog.2023.109415.

[280] C. Liu, L. Shi, X. Zhao, and J. Qiu, “Adaptive matching norm based disparity estimation from light field data”, *Signal Processing*, vol. 209, p. 109 042, 2023. DOI: `https://doi.org/10.1016/j.sigpro.2023.109042`.

[281] Y. Liu, Y. Pan, K. Luo, Y. Liu, and L. Zhang, “Feamnet: Light field depth estimation network based on feature extraction and attention mechanism”, in *2023 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2023, pp. 1–8. DOI: `https://doi.org/10.1109/IJCNN54540.2023.10191959`.

[282] H. Sheng, Y. Liu, J. Yu, *et al.*, “Lfnat 2023 challenge on light field depth estimation: Methods and results”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3472–3484. DOI: `https://doi.org/10.1109/CVPRW59228.2023.00350`.

[283] T. Wang, R. Chen, R. Cong, *et al.*, “Epi-guided cost construction network for light field disparity estimation”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3437–3445. DOI: `https://doi.org/10.1109/CVPRW59228.2023.00346`.

[284] X. Wang, C. Tao, and Z. Zheng, “Occlusion-aware light field depth estimation with view attention”, *Optics and Lasers in Engineering*, vol. 160, p. 107 299, 2023. DOI: `https://doi.org/10.1016/j.optlaseng.2022.107299`.

[285] M. Xiao, C. Lv, and X. Liu, “Fpattnet: A multi-scale feature fusion network with occlusion awareness for depth estimation of light field images”, *Sensors*, vol. 23, no. 17, p. 7480, 2023. DOI: `https://doi.org/10.3390/s23177480`.

[286] D. Yang, Z. Cui, H. Sheng, *et al.*, “An occlusion and noise-aware stereo framework based on light field imaging for robust disparity estimation”, *IEEE Transactions on Computers*, 2023. DOI: `https://doi.org/10.1109/TC.2023.3343098`.

[287] S. Zhang, N. Meng, and E. Y. Lam, “Unsupervised light field depth estimation via multi-view feature matching with occlusion prediction”, *arXiv preprint arXiv:2301.08433*, 2023. DOI: `https://doi.org/10.1109/TCSVT.2023.3305978`.

[288] W. Zhou, L. Lin, Y. Hong, Q. Li, X. Shen, and E. E. Kuruoglu, “Beyond photometric consistency: Geometry-based occlusion-aware unsupervised light field disparity estimation”, *IEEE Transactions on Neural Networks and Learning Systems*, 2023. DOI: `https://doi.org/10.1109/TNNLS.2023.3289056`.

[289] P. Zhou, L. Shi, X. Liu, J. Jin, Y. Zhang, and J. Hou, “Light field depth estimation via stitched epipolar plane images”, *IEEE Transactions on Visualization and Computer Graphics*, 2023. DOI: `https://doi.org/10.1109/TVCG.2023.3344132`.

[290] J. Berent and P. L. Dragotti, “Segmentation of epipolar-plane image volumes with occlusion and disocclusion competition”, in *2006 IEEE Workshop on Multimedia Signal Processing*, IEEE, 2006, pp. 182–185. DOI: `https://doi.org/10.1109/MMSP.2006.285293`.

[291] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields.", *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73–1, 2013. DOI: `https://doi.org/10.1145/2461912.2461926`.

[292] J. Canny, "A computational approach to edge detection", *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986. DOI: `https://doi.org/10.1109/TPAMI.1986.4767851`.

[293] J. Yu, L. McMillan, and S. Gortler, "Surface camera (scam) light field rendering", *International Journal of Image and Graphics*, vol. 4, no. 04, pp. 605–625, 2004. DOI: `https://doi.org/10.1142/S0219467804001567`.

[294] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing", *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009. DOI: `https://doi.org/10.1145/1531326.1531330`.

[295] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs", *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1521–1525, 2004. DOI: `https://doi.org/10.1109/TPAMI.2004.102`.

[296] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang, "Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures", in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 2331–2338. DOI: `https://doi.org/10.1109/CVPR.2006.244`.

[297] I. Gheţa, C. Frese, M. Heizmann, and J. Beyerer, "A new approach for estimating depth by fusing stereo and defocus information", *Informatik 2007–Informatik trifft Logistik–Band 1*, 2007.

[298] P. Favaro, S. Soatto, M. Burger, and S. J. Osher, "Shape from defocus via diffusion", *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 518–531, 2008. DOI: `https://doi.org/10.1109/TPAMI.2007.1175`.

[299] F. Li, J. Sun, J. Wang, and J. Yu, "Dual-focus stereo imaging", *Journal of Electronic Imaging*, vol. 19, no. 4, p. 043 009, 2010. DOI: `https://doi.org/10.1117/1.3500802`.

[300] I. Gheţa, C. Frese, and M. Heizmann, "Fusion of combined stereo and focus series for depth estimation", *INFORMATIK 2006–Informatik für Menschen, Band 1*, 2006.

[301] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241. DOI: `https://doi.org/10.1007/978-3-319-24574-4_28`.

[302] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein, *Stanford lytro light field archive*, 2016. [Online]. Available: `http://lightfields.stanford.edu/LF2016.html`.

[303] L. McMillan and G. Bishop, “Plenoptic modeling: An image-based rendering system”, in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 39–46. DOI: https://doi.org/10.1145/218380.218398.

[304] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, “The lumigraph”, in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 43–54. DOI: https://doi.org/10.1007/978-0-387-31439-6_8.

[305] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, “Plenoptic sampling”, in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 307–318. DOI: https://doi.org/10.1145/344779.344932.

[306] I. Geys, T. P. Koninckx, and L. Van Gool, “Fast interpolated cameras by combining a gpu based plane sweep with a max-flow regularisation algorithm”, in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, IEEE, 2004, pp. 534–541. DOI: https://doi.org/10.1109/TDPVT.2004.1335283.

[307] A. M. Siu and R. W. Lau, “Image registration for image-based rendering”, *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 241–252, 2005. DOI: https://doi.org/10.1109/TIP.2004.840690.

[308] A. Kubota, K. Aizawa, and T. Chen, “Reconstructing dense light field from array of multifocus images for novel view synthesis”, *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 269–279, 2006. DOI: https://doi.org/10.1109/TIP.2006.884938.

[309] T. Georgiev and A. Lumsdaine, “Reducing plenoptic camera artifacts”, in *Computer Graphics Forum*, Wiley Online Library, vol. 29, 2010, pp. 1955–1968. DOI: https://doi.org/10.1111/j.1467-8659.2010.01662.x.

[310] D. J. MacKay, “A practical bayesian framework for backpropagation networks”, *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992. DOI: https://doi.org/10.1162/neco.1992.4.3.448.

[311] R. M. Neal, *Bayesian learning for neural networks.* Springer Science & Business Media, 2012, vol. 118. DOI: https://doi.org/10.1007/978-1-4612-0745-0.

[312] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”, in *international conference on machine learning*, PMLR, 2016, pp. 1050–1059. DOI: https://doi.org/10.48550/arXiv.1506.02142.

[313] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?”, *Advances in neural information processing systems*, vol. 30, 2017. DOI: https://doi.org/10.48550/arXiv.1703.04977.

[314] E. Ilg, O. Cicek, S. Galesso, *et al.*, “Uncertainty estimates and multi-hypotheses networks for optical flow”, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 652–667. DOI: https://doi.org/10.1007/978-3-030-01234-2_40.

[315] L. Ardizzone, J. Kruse, S. Wirkert, *et al.*, “Analyzing inverse problems with invertible neural networks”, *arXiv preprint arXiv:1808.04730*, 2018. doi: https://doi.org/10.48550/arXiv.1808.04730.

[316] L. Ardizzone, J. Kruse, C. Lüth, N. Bracher, C. Rother, and U. Köthe, “Conditional invertible neural networks for diverse image-to-image translation”, in *DAGM German Conference on Pattern Recognition*, Springer, 2020, pp. 373–387. doi: https://doi.org/10.1007/978-3-030-71278-5_27.

[317] S. Im, H.-G. Jeon, S. Lin, and I. S. Kweon, “Dpsnet: End-to-end deep plane sweep stereo”, *arXiv preprint arXiv:1905.00538*, 2019. doi: https://doi.org/10.48550/arXiv.1905.00538.

[318] A. Paszke, S. Gross, F. Massa, *et al.*, “Pytorch: An imperative style, high-performance deep learning library”, *Advances in neural information processing systems*, vol. 32, 2019. doi: https://doi.org/10.48550/arXiv.1912.01703.

[319] A. P. Dawid and P. Sebastiani, “Coherent dispersion criteria for optimal experimental design”, *Annals of Statistics*, pp. 65–81, 1999. doi: https://doi.org/10.1214/aos/1018031101.

[320] Y. Ovadia, E. Fertig, J. Ren, *et al.*, “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift”, *Advances in neural information processing systems*, vol. 32, 2019. doi: https://doi.org/10.48550/arXiv.1906.02530.

[321] K. Hara, D. Saitoh, and H. Shouno, “Analysis of dropout learning regarded as ensemble learning”, in *International Conference on Artificial Neural Networks*, Springer, 2016, pp. 72–79. doi: https://doi.org/10.1007/978-3-319-44781-0_9.

[322] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks”, in *International Conference on Machine Learning*, PMLR, 2017, pp. 1321–1330. doi: https://doi.org/10.48550/arXiv.1706.04599.