Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Dense Light Field Reconstruction: from Depth-based to Learning-based Approaches

Yang Chen

October, 2021

A dissertation submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism ‘Ready Steady Write’, located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: ____________________ Date: ____________________

Abstract

Light field imaging and processing is an emerging technique that motivates production of 3D visual content and makes it possible to provide high quality immersive 3D experiences. The principle of light fields is designed to describe all light rays passing through a given volume in 3D space, and only suitable acquisition can construct a dense light field, which is advantageous in practical applications, such as medical imaging, computer animation and post-capture photography. However, limited by the processing capability, acquiring a sufficient amount of high dimensional information usually leads to significant computational complexity and inaccuracy. Thus, to bridge the gap between acquisition and the required visual information, in this thesis, we are looking for the establishment of an efficient and accurate light field reconstruction framework, which only requires sparse light field input.

First, we will introduce our contribution to depth estimation from the 4D light fields and its application to render novel views for light field reconstruction. We build an optical flow framework to estimate disparity by tracking pixel movement. To further improve the efficiency, instead of using traditional global optimization, we use an alternative edge-aware filtering to efficiently encourage the smoothness while retaining high-frequency information. Compared to other state-of-the-art methods, our framework is capable of extracting geometrical information in an efficient and accurate fashion. Furthermore, we also move to light field reconstruction by warping input views to novel locations with the estimated disparity map.

Second, we investigate subsampling and reconstruction strategies for light fields processing. Limited angular resolution of acquired light fields is one of the issues in light field data, which usually is massive requiring high computational expense to process. Different from numerous previous works focusing on employing novel techniques, we chose an unique angle to optimize the performance of light field reconstruction, which is concentrated on comparing various commonly used view selection strategies. This work could benefit a wide range of applications, such as camera hardware design, light field compression and light field rendering.

Last but not least, we propose a deep learning based framework for light field view synthesis. With the booming development of data-driven techniques, deep learning based methods have been successfully applied to light field related tasks, such as material recognition, depth estimation and view synthesis. However, learning methods usually require a huge amount of data and collecting sufficient light field data is a challenging task due to expensive acquisition. Thus, we employ cycle consistency to the light field view synthesis task, which enables training to be performed in a self-supervised manner and avoids the requirement for huge training data. Experimental results show that our

method outperforms other state-of-the-art light field view synthesis methods, especially when input views have wider disparity.

Acknowledgements

I feel tremendously lucky to have received a lot of help, support and encouragement from many people during these years, without which this thesis would not have been accomplished.

I first wish to thank Prof. Aljosa Smolic and Dr. Martin Alain for their support from beginning to end of my PhD journey. Aljosa made a great effort on supervising me during the last four years, from theoretical derivation to the wording in professional publications. He always is the lighthouse for and guiding me on the right path. Martin works closely as postdoc and co-author with me in developing light field processing methods, generously sharing profound light field knowledge and refining the research papers. I also would like to thank Gail, our dearest administrator and representative of Irish kindness, who helps me deal with all official procedures. Moreover, I am very grateful to have Prof. Atanas Gotchev (Tampere University) and Prof. Michael Manzke (Trinity College Dublin) as examiners for thesis correction, along with Prof. Donal O'Mahony (Trinity College Dublin) as chair for oral viva.

Appreciation also would be given to my PhD peers: Sebastian Lutz, Tejo Chalasani, Rafael Monroy, Koustav Ghosal, David Hardman, Pierre Matysiak, Ailbhe Gill, Matthew Moynihan and Colm O'Fearghail. We spent so much joyful time together and have numerous highly intellectual talk. I appreciate their help for getting familiar with local culture and the potential diversity within this boundless world. I am grateful for being offered this PhD position which gives me a chance to be a member of the V-SENSE team and to pursue the dream of being a computer scientist. I wish to thank all those nice people for creating such the warm work environment together. Additionally, Valve has some credits on this list as it releases me from academic pressure and brings me many happy memories at spare time.

I must present my deepest gratitude to all my family members for their life-long love, support and encouragement. My grandparents, my parents, my uncle and aunts, my siblings and my 2-years-old nephew, all-of-them give me unconditional support and motivation to fight and conquer. This dissertation is dedicated to them.

I want thank Science Foundation Ireland who funds my complete PhD study. I also want to thank Nvidia generously sponsors me with a powerful GPU which equips me as a qualified data scientist.

2020 is destined to be an unforgettable year, which brings so much rapid changes to the whole world. I wish every reader of this thesis could stay healthy and happy as forever as possible.

For My Dearest Family

Contents

List of Figures

List of Tables

Nomenclature

DSLR camera	**D**igital **S**ingle-**L**ens **R**eflex camera
EPI	**E**pipolar **P**lane **I**mage
CNN	**C**onvolutional **N**eural **N**etwork
MPI	**M**ulti**P**lane **I**mage
MSI	**M**ulti**S**phere **I**mage
LCD	**L**iquid-**C**rystal **D**isplay
FF	**F**eature **F**low
PF	**P**ermeability **F**ilter
CPM	**C**oarse-to-fine **P**atch **M**atch
NNF	**N**earest **N**eighbor **F**ield
WLS	**W**eighted **L**east **S**quares
BOOM	**B**inarized **O**ctal **O**rientation **M**aps
SABOM	**S**patio-**A**ngular **B**inarised **O**rientation **M**aps
SIFT	**S**cale-**I**nvariant **F**eature **T**ransform
CPM_FF	**C**oarse-to-fine **P**atch **M**atch_**F**eature **F**low
CPM_PF	**C**oarse-to-fine **P**atch **M**atch_**P**ermeability **F**ilter
DIBR	**D**epth **I**mage **B**ased **R**endering
MSE	**M**ean **S**quare **E**rror
SOTA	**S**tate-**O**f-**T**he-**A**rts
IR	**I**nput **R**atio
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
SSIM	**S**tructural **S**imilarity **I**ndex **M**easure
H-V	**H**orizontal-**V**ertical
V-H	**V**ertical-**H**orizontal

1 Introduction

In this chapter, we first explain the motivation for exploring reconstruction of dense lights field by presenting our general research statement. Then, the structure of the thesis and the main content of each following chapter are presented. At the end of this chapter, the main contributions are highlighted and related publications are listed respectively.

1.1 Motivation

We live in the digital era. From advanced professional monitor systems to personal smartphones and computers, modern graphic display systems have become an indispensable part of our work and life. These display systems aim to visualize information to the human visual perception system by reconstructing visual content from acquired or received data. The visual perception system of human beings can easily understand the 3D structure; however, nowadays, most of the digital systems are only capable of displaying 2D (or pseudo 3D) content. It is a long-lasting desire to bring our digital displays to the age of actual 3D content. Various scientific or practical attempts, including light fields, super multi-view, integral imaging and holographic displays, have been made towards recreating the 3D scene in digital devices. This 3D reconstruction problem, which refers to a long-standing research area in computer vision and graphics, has inspired numerous applications in gaming, medicine and film industry.

Many 3D reconstruction methods have been developed to be applied to a variety of data, each of which is gathered through different acquisition processes, which usually is composed of complex and multiple sensors. A sufficient amount of high-quality data is the foundation stone to depict the desired real-world 3D scene accurately. However, comprehensive data collection of the detailed 3D scene is expensive regarding time and finances, which is mostly limited by the hardware capability. **Hence, a robust and efficient software framework becomes the practical solution to close the gap between the collected data from available sensors that are typically limited and the requirement of high-quality visual content for the future generation**

of display systems.

The 4D light field is a typical technique utilized for processing high dimensional visual information. It was originally introduced by Levoy et al. [8] for the purpose of image-based rendering. The generalized framework of the 4D light field is designed to reconstruct visual information of a target scene by mapping multi-perspective images to a multi-dimensional light field function. The output light field data can be represented as the collection of images arranged on a 2D grid with slight horizontal and vertical perspective shifts. Due to the potential capability to describe the 3D representation of the scene, light fields have gained a lot of attention in both industry and academia in recent years and enabled various practical applications in extending dimensions beyond 2D spaces, such as refocusing of photography and virtual reality. Light field images contain rich 3D information because they provide coherent information of every pixel not only along the spatial dimensions but also in additional angular dimensions. However, existing light field acquisition systems, varying from gantry controlled DSLR cameras, through arrays of DSLR cameras, to the recently developed consumer-level plenoptic cameras, can only produce sampled representations of light fields as collections of images of the target scene, and generally suffer from a spatio-angular resolution trade-off as a result of the limited hardware design.

1.2 Structure of the Thesis

In this thesis, we will concentrate on the reconstruction of high quality dense light fields given a limited amount of multi-perspective views. We investigate two approaches to solve this problem: the first approach is a traditional disparity-based model and the second approach employs the deep learning technique. Besides, high-level light field selection strategies are presented to optimize the performance of light field reconstruction.

In Chapter 2, we first introduce the development of light field imaging, along with the derivation of the corresponding theoretical framework and development of its practical implementation. We also review comprehensive literature regarding previous light field reconstruction work, including depth estimation from light fields and depth-based image rendering.

In Chapter 3, we present our initial attempt to synthesize views based on the depth-based image rendering. This method follows the typical research path relying on the geometric information of the target scene, e.g. depth information, as the prior knowledge. More specifically, the captured sparse light field data is processed to estimate the geometric primitives. These primitives are utilized to construct the underlying structure of the

3D scene and estimate expected views under specific circumstances. The advantage of this type of method is that the global structure of the 3D scene can be established and corresponding visual information can be continuously obtained by rendered from the model.

In Chapter 4, high-level view selection strategies are discussed to optimize the performance of the light field reconstruction. Although state-of-the-art and our previous work achieves impressive performance on the task of light field view synthesis, limited attention has been paid to explore the optimal view selection strategy, which has potential impact on the performance of light field reconstruction. We selected one benchmark method and utilized it to investigate various subsampling and reconstruction strategies, and then our experiments demonstrate the optimal strategies in each case.

In Chapter 5, a data-driven approach is described to explore the possibility to learn a model for reconstruction from limited input training samples. With the boom of deep learning techniques, prior work showed remarkable performance and capability to generalize to tasks related to light field reconstruction. As the training process usually requires the support of a huge amount of data, which is the one shortcoming in the area of light field processing, it is reasonable to transfer the comprehensive prior knowledge from numerous single image and video datasets to the light field domain by establishing the connection of the light field view synthesis problem to a video interpolation problem. Furthermore, we apply cycle consistency to enable training in a self-supervised manner, which also reduces the required amount of light field training data. Our experiments demonstrate the success of our method, which outperforms state-of-the-art methods on various benchmark datasets.

In Chapter 6, we conclude our main contributions introduced in this thesis and discuss the potential directions to be explored in future work.

1.3 Main Contributions and Publications

The following contributions regarding light field reconstruction are results of my PhD research, and the corresponding publications are listed as well:

In Chapter 3, we propose a spatial-angular edge-aware optical flow-based framework and apply it along the angular domain light fields to estimate the corresponding depth maps.

- [9] Yang Chen, Martin Alain, and Aljosa Smolic, *Fast and Accurate Optical Flow-based Depth Map Estimation from Light Fields,* in Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP), IPRCS, Aug 2017. (**Best Paper Award**)

In Chapter 4, we discuss view selection strategies for light field subsampling and reconstruction, and experimentally evaluate the performance of each strategy.

- [10] Yang Chen, Martin Alain, and Aljosa Smolic, *A Study of Efficient Light Field Subsampling and Reconstruction Strategies,* in Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP), IPRCS, Aug 2020.

In Chapter 5, we expand a video interpolation method in the angular dimension to be applied to light field view synthesis. This involves cycle consistency, which allows the network to be trained in a self-supervised manner.

- [11] Yang Chen, Martin Alain, and Aljosa Smolic, *Self-supervised Light Field View Synthesis Using Cycle Consistency,* in 22th International Workshop on Multimedia Signal Processing (MMSP), IEEE, Sep 2020. (nominated for Best Paper Award)

2 Background

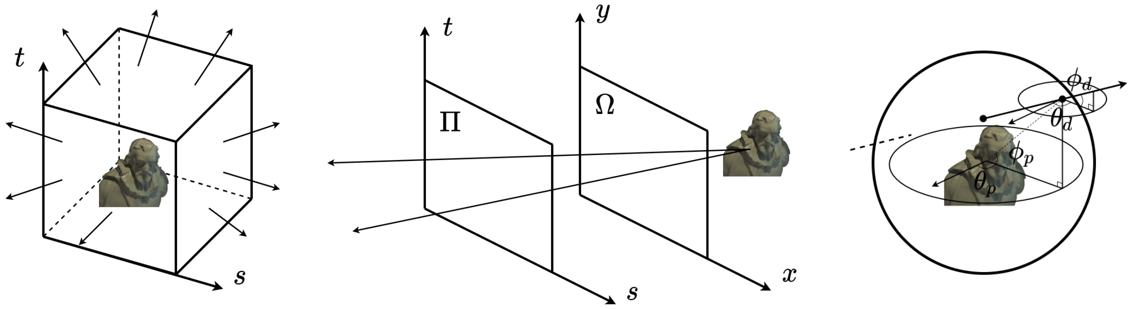
In this chapter, we will present the fundamental knowledge that establishes the basis for developing the contributions presented later in this thesis. We will first introduce the background concepts about light field imaging, *i.e.* light field parametrization, as well as existing various light field acquisition methodologies. Then, depth map estimation from light field is introduced as the initial research problem. We also will talk about the resolution issue of collected light field data which is brought by the limitations of existing acquisition systems. We also review state-of-the-art related work of depth estimation and light field reconstruction, respectively.

2.1 Light Field Parametrization, Acquisition and Application

After decades of development of the light field technology, the application of light field on classic computer vision problems have received considerable attention both from academia and industry. 4D light field images provide rich scene information via capturing all light rays within a given volume of space. Compared to conventional photography, not only two spatial dimensions but also two extra angular dimensions are provided. In particular, this additional information inherently reveals more details about the geometry of the targeted scene. This boosts multiple applications, such as light-field microscopy [12], post-capture refocusing [13], image-based rendering [14], near-eye light field displays [15], glasses-free 3D display [16] and more. In this thesis, we will illustrate how this rich information is processed to benefit vision tasks. Here, we first introduce the derivation of the light field theory. We refer readers to [17] for a review about light field processing.

2.1.1 Early Theories

According to the best of our knowledge, the very first article describing the idea of capturing light-field-like data to create 3D photographs was first proposed by Lippmann



(a) Cube

(b) Two-plane

(c) Sphere

Figure 2.1: Illustration of different parameterization of light rays.

in 1908 [18], although the original term they used is "Photographies Integrales" ("Integral Photographs"). After almost one-century development of computational imaging techniques, Adelson and Bergen firstly proposed to formalize the **light**, physical medium transferring visual information, as the *plenoptic function* concept in 1991 [19], which is expressed as a 7D function:

$$P = P(\theta, \phi, \lambda, t, V_x, V_y, V_z) \tag{2.1}$$

This function describes the light from a scene as the intensity of all the rays passing through the observation position at the location (V_x, V_y, V_z) in the 3D spaces at the direction (θ, ϕ) in the spherical coordinate system, the wavelength λ and the time t.

As one of the subsequent work, in 1995, McMillan and Bishop [20] proposed to simplify the 7D plenoptic function to 5D by employing a fixed time and wavelength constraint, *i.e.* for only primary colors, and describing the flow of the light at different 3D locations along with various 2D directions:

$$P = P(\theta, \phi, V_x, V_y, V_z) \tag{2.2}$$

2.1.2 4D Light Field and Various Representation

To further reduce the complexity of computation and meet the requirement of practical applications, the 4D light field is proposed in [14, 21], which reduced 5D parametrization to the 4D space. This simplification is based on an important assumption that the radiance does not change along a line unless blocked, so the 4D function can be expressed in this way:

$$P = P(\theta, V_x, V_y, V_z) \tag{2.3}$$

The core idea is to apply one-directional constraint and only consider the light passing through the given direction instead of any direction. The beforehand representation can be instantiated as the rays from the target scene through a closed convex hull, *i.e.* virtual cube box, as shown in the Figure 2.1a, of which six surface planes are parametrized by a pair of orthogonal coordinates s and t, which could define one plane (s, t). To exclusively determine and describe the ray coming from the scene, the direction of any ray is parametrized by a double-plane system, including plane (s, t) and another parallel plane (x, y), as shown in the Figure 2.1b. Any directional ray can be mapped to a 4D coordinate (s, t, x, y), which derives the commonly known two-plane representation. Levoy and Hanrahan call this each pair of planes a *light slab* [14]; Gortler et al. [21] propose an arrangement of six pairs of planes called the *lumigraph*. As this two-plane representation of the light rays provides a more effective way to describe rays of light, Levoy et al. [8] summarized 4D light field utilizing this representation and observed the beginning of the blooming era of the light fields research with computational photography and computer vision techniques.

Two-plane Representation

In this thesis, we will use the two-plane representation to formalize the geometric structure of the light field, which is the most utilized representation in related researches nowadays. Mathematically, as visualized in the left part of Figure 2.2, the two-plane parametrized light field can be represented as a 4D function:

$$L : \Omega \times \Pi \to \mathbb{R}, (s, t, x, y) \to L(s, t, x, y) \tag{2.4}$$

in which the image plane Ω represents the spatial distribution of light rays, indexed by (x, y), while the focal plane Π corresponds to their angular distribution, indexed by (s, t). To visualize a 4D light field image, perhaps the easiest way is to consider it as a collection of views, also called sub-aperture images, taken from several viewpoints arranged on a 2D grid with slight horizontal and vertical perspective shift. The light field can then be considered as a matrix of views (see the middle part of Figure 2.2). Note that there is an assumption in this thesis when using such representation that the different views are rectified along the same angular dimension.

Epipolar Plane Image. Another common way to represent light fields is as a collection of epipolar plane images (EPIs). The concept of epipolar plane image is originally from the stereo geometry research and introduced to as epipolar geometry [22]. Later, the epipolar geometry is generalized for the case of multiple images by Bolles et al. [23]. In the case of light fields, an epipolar plane image is a 2D slice of the full 4D light field obtained by fixing one spatial and one angular dimension. For example, by fixing the

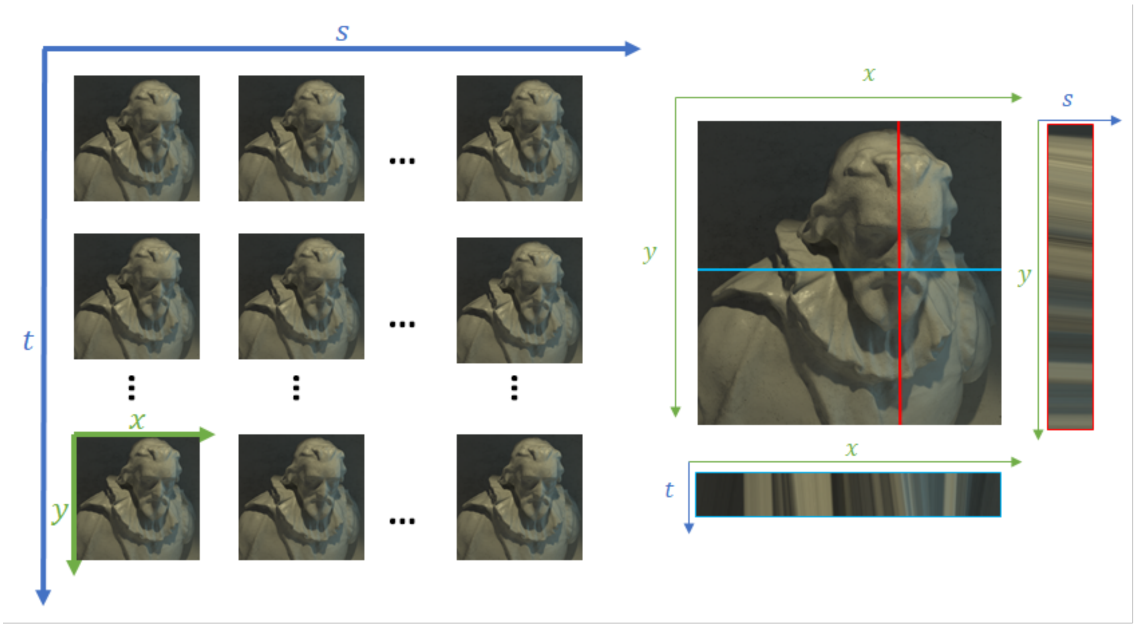


Figure 2.2: Light field two-plane parametrization and epipolar plane images representation.

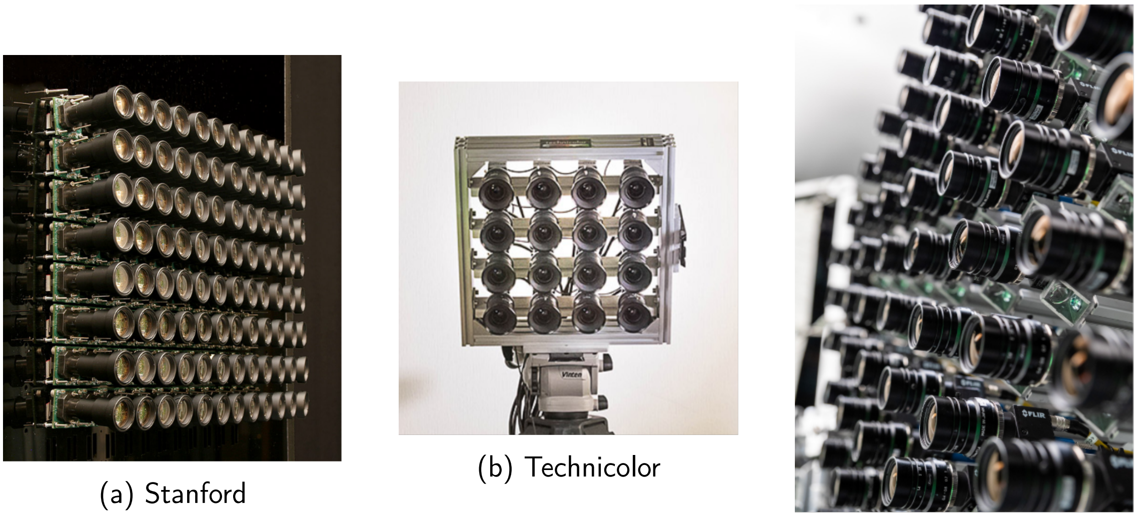
xs-plane or *yt*-plane:

$$L_{t^*,y^*}(s,x) \rightarrow L(s,t^*,x,y^*) \tag{2.5}$$
$$L_{s^*,x^*}(t,y) \rightarrow L(s^*,t,x^*,y) \tag{2.6}$$

These two ways of extracting epipolar plane images correspond to horizontal and vertical directions respectively, as shown in the right part of Figure 2.2. Such representation can help visualising the angular information using only 2D images, including the implicit expression of the depth of the scene.

Spherical Representation

While we adopt the 2 parallel plane parameterization in the rest of the thesis, other parameterization exists for various scenario. One example is the spherical representations, which is especially suited for immersive virtual reality applications. The spherical representation of the light field is designed to avoid the artefact on the boundary of the imaging plane, which usually happens with the two-plane representation, as it is expected to acquire the light field in a (nearly) uniform fashion. The 4D spherical parameterization was introduced for light field acquisition systems that are set at a fixed position while moving along a 360-degree direction, as shown in 2.1c, which are useful for applications such as virtual reality. Ihm et al. [24] proposed the spherical representation of the light rays that can be expressed as the combination of two functions f_d and



(a) Stanford

(b) Technicolor

(c) SAUCE

Figure 2.3: Light field acquisition systems - camera arrays.

f_p:

$$
\begin{aligned}
L^{sphere} &= f_d(\theta_d, \phi_d) && (2.7)\\
&= (f_p(\theta_p, \phi_p))(\theta_d, \phi_d) && (2.8)
\end{aligned}
$$

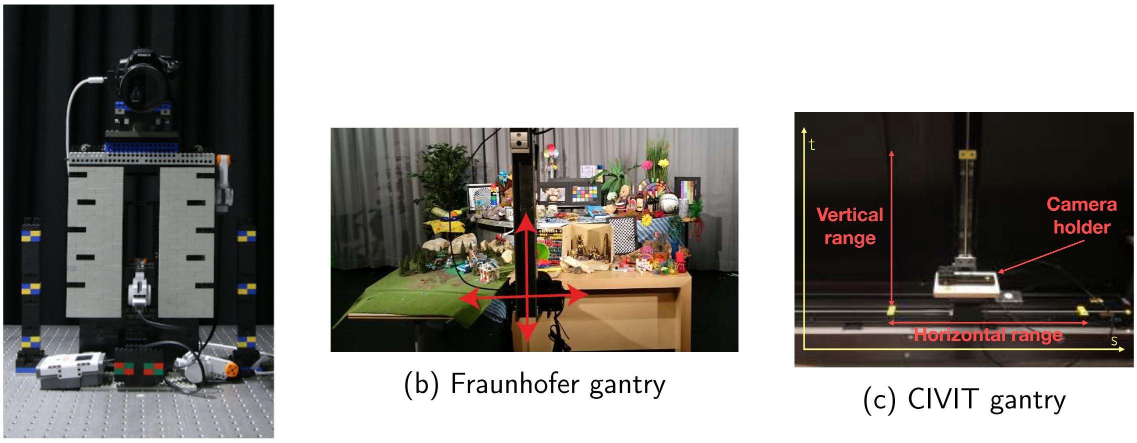
where f_p is a function defined on a sphere whose value is function f_d. The term (θ_p, ϕ_p) indicates the interaction point between the light ray with the positional sphere, and (θ_d, ϕ_d) determines the orientation of the light ray at the point (θ_p, ϕ_p) (see Figure 2.1c).

Other two sphere-based parameterization, two-sphere parametrization (2SP) and sphere-plane representation (SPP), are introduced by Camahort [25]. Each ray is parameterized by its intersection points with the same sphere (2SP), or by its angle and the 2D coordinate of the intersection point of the ray and the orthogonal plane (SPP).

2.1.3 Acquisition Systems

Camera Array

Camera array is a typical imaging system that is capable of capturing sub-aperture views with a single exposure of all cameras at the same time. In such a way, each sub-aperture view from the two-plane parametrizations of light field can be directly captured by cameras located at each perspective in the array. The early-stage effort for the camera array-based system is from the Stanford group [26], who build a camera array consists



(a) Stanford lego gantry

(b) Fraunhofer gantry

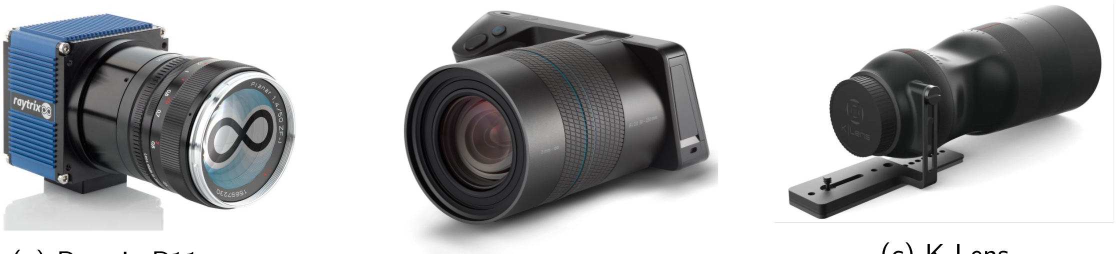
(c) CIVIT gantry

Figure 2.4: Light field acquisition systems - gantry controlled camera.

of more than one hundred cameras arranged in the same plane, shown as Figure 2.3a. A more recent camera array system is proposed by Sabater et al. [27] with sixteen DSLR cameras, as shown in Figure 2.3b, which extended the capability to capture light field video. Along with the hardware system, a post-processing pipeline is also introduced by [27], including color correction, calibration, depth estimation and view rendering. Another multi-camera array system is constructed by Herfet et al. [28] producing the light field video with high definition, as shown in Figure 2.3c. This work targets to enhance the quality of light field content to match the up to date high-performance display devices. The potential of camera rigs such as [28] for industrial film-making and post-production is explored by Trottnow et al. in a follow-up work [29]. Although camera array-based systems are efficient to capture light field, the density of captured light field is limited by the number of cameras that can be set up, and it might require further post-processing to reconstruct the dense light field. In addition, these systems are bulky and inconvenient to carry around; thus, its potential application scenarios are restricted.

Gantry Controlled Camera

One alternative method to acquire light field is to use a gantry controlled camera moving through multiple perspectives as proposed in [30], as shown in Figure 2.4a. Ziegler et al. utilize a gantry controlled DSLR camera to capture large scenes [5], which is guided by a precise linear axe system, as shown in Figure 2.4b. Moreschini et al. released a dense captured with a horizontally motorized positioning system [7], as shown in Figure 2.4c, which can used as ground truth for light field reconstruction challenge [31]. With the gantry controlled camera, it is easier to move and set up the system under various circumstances. The spatio-angular quality of the light field can be determined by the



(a) Raytrix R11 camera

(b) Lytro camera

(c) K-Lens

Figure 2.5: Light field acquisition systems - handheld camera.

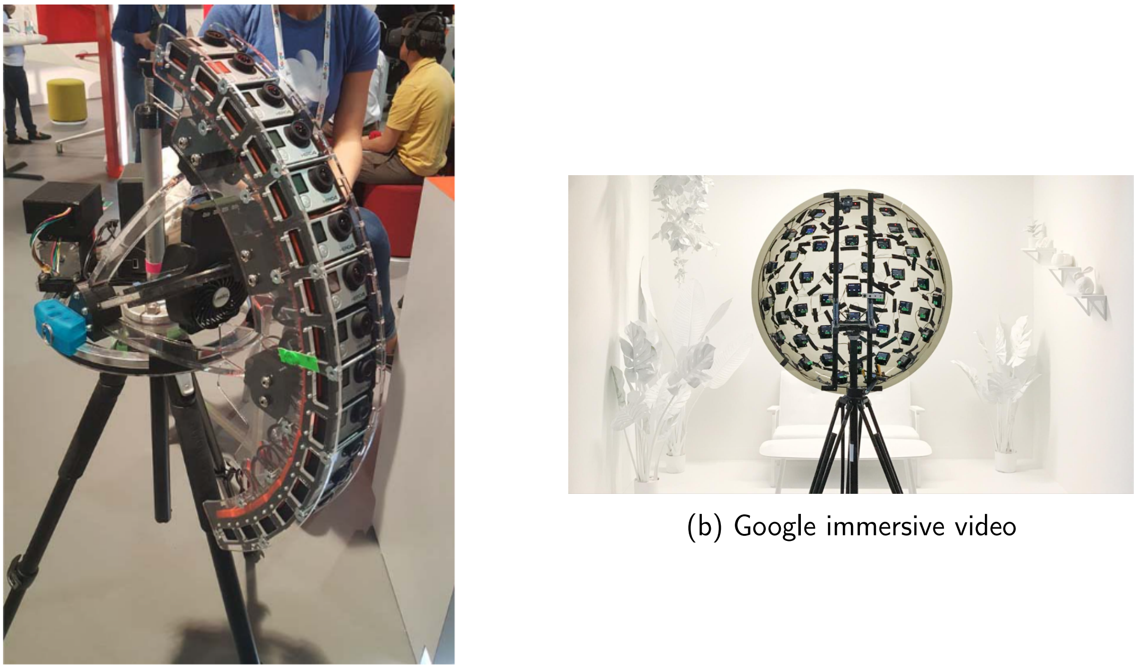
specification of the single camera and the accuracy of the gantry control system. The main shortcoming of this acquisition method is that it is time expensive to complete the camera movement. Furthermore, since it requires some time to complete the imaging, it can only be used to capture light field for a static scene. Note that the gantry setup can also be modified to capture spherical light fields, as introduced by [30].

Handheld light field camera

As opposed to camera arrays or gantry controlled camera systems described above, which are mainly restricted to lab setups, handheld devices have been designed allowing for more portable light field capture. These devices either rely on micro-lens technology or add-on inter-reflection lenses.

As the core component of a micro-lens array camera, also called plenoptic cameras, micro-lens arrays are formed as multiple micro-lenses arranged in an grid, and placed in front of the image sensor to form a two-plane system inside one camera. While in traditional 2D cameras the main lens integrates light rays from different directions onto a single point on the sensor, the micro-lenses redistribute the rays coming from different angles to different sensor locations. Raytrix [32] markets commercial plenoptic cameras and offers their signature plenoptic camera R11 with a three-megapixel effective resolution, as shown in Figure 2.5a. The consumer-level plenoptic camera Lytro Illum was then introduced to the public by Lytro Inc. [33], as shown in Figure 2.5b. The output raw data from existing plenoptic cameras require post-processing, including decoding, calibration, rectification and color correction [34, 35], to obtain sub-aperture views of the light field. A similar structure of the plenoptic camera also been adapted to some systems for specific tasks, such as light field microscopy for medical inspection.

Aside from plenoptic cameras, an inter-reflection camera add-on was introduced by Manakov et al. [36], later developped in a product by the K-lens company, as shown in Figure 2.5c. This optical component can be directly attached to a DSLR camera and



(b) Google immersive video

(a) Google VR

Figure 2.6: Light field acquisition systems - immersive video systems.

then performs the light field acquisition. This system is designed to exploits the inter-reflections mechanism in a kaleidoscopic structure implemented with parallel mirrors as image multiplier to sample the aperture views of light fields.

Although these micro-lens array or inter-reflection based plenoptic devices are easy to carry and go, the quality, especially the spatial resolution, of the output image is still not competitive to images from cutting edge DSLR cameras. This is due to the fact that the full spatio-angular information has to be multiplexed onto a single sensor. In addition, color inconsistencies and noise can occur in the sub-aperture images due to optics limitations.

Immersive Video Acquisition

Recently, Google released a spherical light field acquisition system demonstrating the potential capability of light field techniques for virtual reality applications [37]. This rotating system similar to gantry-based systems desceibed above, as shown in Figure 2.6a, consists of sixteen consecutive GoPro cameras arranged on semi-circle, which is able to capture a static scene in 360 degrees from the target point of view. Then, high quality immersive panoramic light fields can be rendered within a consumer-level virtual reality display. They generally follow a more conventional light field rendering pipeline, which utilizes depth geometrical primitives to synthesize continuous views in real-time. Later, Broxton et al. [38] introduced another light field video acquisition system, which has a hemispherical arrangement of 46 low-cost action cameras and only covers a limited portion of the full 360 field of view, as shown in Figure 2.6b, but is capable of recording six-degree-of-freedom (6 DOF) video content. A specific layered mesh representation is designed for a better quality of the final panoramic light field content.

2.1.4 Applications of Light Fields

Light Field Imaging for Medical Inspection

Light field technology has been used to improve medical imaging inspection, especially microscopy. A light field microscope was first introduced by Levoy et al. [12] as a 3D computational imaging approach to capture all light passing through a specimen. A prototype of this system is implemented by inserting a micro-lens array to a conventional microscope, which enables computationally synthesizing focal stacks, flexibly controlling depth of field and achieving full volumetric reconstruction. A 3D deconvolution microscope is proposed by [39] which provide improved quality of spatial resolutions compared to [12]. A camera array-based light field microscope is introduced by Lin et al. [40], which contains a two-stage relay system on the aperture-plane. Prevedel et al. present the potential of a light field microscope to push the frontier of medical research [41].

Another potential application of light field imaging used for medical inspection is light field microendoscopy. Surgical microendoscopes is a sophisticated inspection device designed especially for minimally invasive operations, so it requires a smaller size of optical components than microscopes. Kwan et al. introduce the proof-of-concept light field microendoscopes but limited by the physical size of available optical components meeting the size constraints of the microendoscopes, the actual product is not produced in this work [42]. Later, Orth et al. implement the concept of light field microendoscopes with the optical fiber bundles. [43]. They utilize the physical property within the fiber bundle to capture spatial and angular light information and construct an ultra-slim light field imaging probe that enables light field imaging inside patients body with a minimized injury.

Post-capture Refocusing Photography

Synthetic depth-of-field effects rendering has been a long-standing research topic in computer graphics and computational photography [44, 45, 46, 47, 48]. Synthetic aperture imaging, also known as refocusing, has been a well known application of light fields since the early stages of research on the topic [49, 50]. Levoy et al. notably introduced the popular "shift-and-sum" refocusing algorithm which consists in shifting the light field views by a disparity value corresponding to the desired focus plane, and averaging the shifted images [50]. Refocusing was revisited by Ng et al. who introduced the concept of Fourier slice photography [51], where the shifting and summing operations are preformed in the Fourier transform domain. Refocusing from light field was also extended to synthesize tilt-shift photography, where the focal plane is *not* fronto-parallel to the camera plane [52, 53]. With the booming of the deep learning techniques, a CNN based refocusing method was also recently proposed by Dayan et al., which only requires a sparse light field as input [54]. Note that sparse light fields create angular aliasing artefacts in defocused blur area of refocused images, thus, high quality light field reconstruction method becomes critical for this application.

Light Field Free Viewpoint Rendering

The first application of light fields was the ability to render images corresponding to novel free viewpoints, without the need for any geometric information [14, 21]. Using the two-plane light field parameterization, novel views are generated by projecting rays from the target viewpoint onto the light field two planes and performing a quadrilinear interpolation, consisting of one bilinear interpolation on the image plane and another bilinear interpolation on the camera plane. Free viewpoint rendering was later combined with refocusing by Isaksen et al. [49]. While the field of view for two-plane light field is limited to the volume of space comprised in-between the two planes, freeviewpoint rendering can also be used with spherical light fields for a full 360 field of view [55]. More recently, the concept of multiplane images (MPIs) [56] and multisphere images (MSIs) [57] have been introduced. These representations can be generated from two-plane and spherical light fields respectively. An MPI consists in a collection of parallel planes covering the depth range of the scene, where each plane is equipped with an RGB texture and alpha map corresponding to the scene objects intersecting the planes. Novel images can be efficiently rendered by warping the texture of each plane to the target free viewpoint using homographies, and compositing the textures from back to front using the "over" operator. A similar principle can be used to render images from MSIs, except that the warping can not be performed using homographies. General light field rendering techniques requires comprehensive information from dense light field to avoid angular aliasing, which motivates the development of light field reconstruction algorithms.

Light Field for Media Production

In recent years. the potential of light field for vidual media production has also been explored [29, 58]. In [29], a complete pipeline of light field video production is summarized, from light field video acquisition using an array of 64 cameras (shown in Figure 2.3c), to calibration and post-processing including depth estimation, point cloud creation, and synthetic aperture rendering, to vizualisation and integration in standard post-production pipeline. In addition, production of immersive content for virtual and augmented reality is another future direction of light field application, as aforementioned in subsection 2.1.3.

Light Field Display

Virtual and augmented reality are among the most trending research areas and usually require a wearable head-mounted display to maximize the immersive experience. Huang et al. introduce a near-eye light field display combining stereoscopic display principles with emerging factored light field [15]. As opposed to main-streaming virtual reality headsets available on the market (such as HTC Vive and Facebook Oculus), this virtual reality display is capable of presenting light fields in an immersive manner with adjustable focus plane. Later, Sluka et al. introduce a new light field headset prototype "CREAL3D", which emits several smaller pinhole-aperture light fields instead of one single light field, guided to the eye by the reflection of mirrors [59]. Furthermore, Avegant introduces a head-mounted display for light field augmented reality applications, which shares a similar outlook to the Microsoft Hololens [60].

Glasses-free 3D display is another development direction of light field displays. Different from a wearable head-mounted display, glasses-free displays aim to reproduce vivid 3D visible content on a digital screen observable for multiple people at the same time. Wetzstein et al. proposed a 3D glasses-free compressive light field display based on a light field. They comprise stacks of light-attenuating LCD layers as a tensor display illuminated by uniform or directional back-lightning [16]. Fattal et al. introduce a multi-directional back-lighting technique which can be used for multiview 3D display on mobile devices [61]. The team from the Looking Glass Factory introduced a desktop holographic display for any professional or hobbyist creating content in 3D [62]. Hirsch et al. present a novel way of displaying a 3D effect on a surface [63]. Instead of utilizing a digital screen, they adapt computational display methods to the creation of hologram-like 3D images on standard print material. More recently, Sony released a holographic display called spatial reality display, which is equipped with eye-tracking and lenticular lenses for glasses-free 3D viewing [64].

2.2 Depth Estimation from Light Field

Depth map estimation is a fundamental research topic in computer vision and graphics, and new approaches have recently emerged taking advantage of light fields. This new imaging modality captures much more information about the angular direction of light rays compared to common approaches based on stereoscopic images or multi-view. Accurate depth map estimation from light fields can benefit a wide range of consecutive

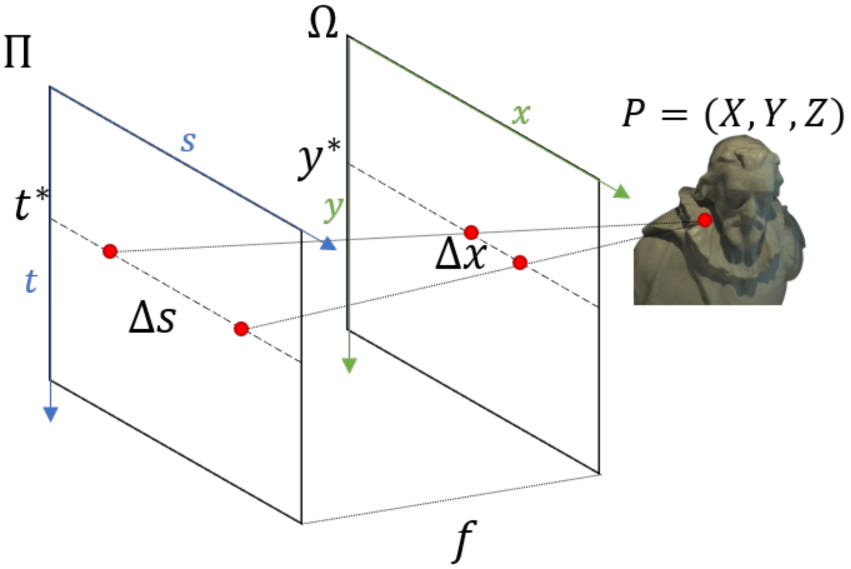


Figure 2.7: Depth estimation from light fields.

applications such as depth-based rendering, digital refocusing and immersive video capturing. As shown in Figure 2.7, the depth Z of the target point P in 3D space can be expressed as:

$$Z = -f\frac{\Delta s}{\Delta x} \tag{2.9}$$

where f is the focal length of the camera and Δs are the baseline distance between camera positions, which are known parameters at the time of capture. In other words, the depth Z can thus be obtained by estimating the disparity Δx, as long as the camera configuration is available. Note that in this thesis we will later focus on disparity estimation, as disparity can also be directly used for light field reconstruction, and does not require any knowledge of the camera parameters. Please note in this thesis, the camera parameters (focal length f and baseline distance Δs) of some light fields are not available, and we are focusing on rectified light fields, on which estimating disparity Δx is sufficient to reconstruct light field subsequently.

2.2.1 Optimization Model Related Methods

Optimization model generally attempts to estimate the depth information by optimizing the various cost functions among the light field structure. Multiple methods taking advantage of the existing literature in stereo disparity estimation have then been proposed to estimate depth from light fields. These techniques rely on various matching approaches to estimate the disparity between views from the light field and a reference view (often the center one). In [65], the authors proposed an accurate block-matching method reaching sub-pixel accuracy based on the Fourier phase-shift theorem with graph cut multi-label optimization. In [66], the authors proposed a complete pipeline using a demultiplexing method to decode views from raw light field data captured with a lenslet camera without demosaicking and then estimate the disparity with a robust block-matching processing [67]. In [68], sparse and accurate matchings are first found, and then interpolated using optical flow. To reduce the complexity, a multi-resolution approach was proposed in [69]. In [70], a global matching is obtained through a low-rank decomposition of the views, refined with homographies.

To better take into account the light field 4D structure, extensions of the previous methods have been proposed based on the analysis of texture patches sampled along the angular dimensions instead of the spatial dimensions. These angular patches, also called SCam, were first exploited in [71]. This work was further extended in [72] to be robust to occlusion. More recently, this idea was included in a global optimization framework [73] in order to obtain a dense depth map estimation.

Several techniques also exploit the light field structure through epipolar plane image, as in such images the slope of a line has a linear relationship with the depth. In [74], depth from high spatio-angular resolution light fields is obtained by first estimating high confidence depth values on the epipolar plane image edges with a sparse representation, and then propagating this information to homogeneous regions using a fine-to-coarse refinement approach. A similar idea about high confidence regions is also introduced by [75], which combines defocus and correspondence cues obtained from the epipolar plane image by shearing the epipolar lines. The shearing angle optimizing the multiple cues response gives the slope of the epipolar lines, and thus the depth. Wanner et al. [76] proposed to estimate the slope of the epipolar lines and its confidence using the structure tensor. In [77] a novel spinning parallelogram operator is introduced and integrated into a depth map estimation framework which advantageously handles noise and occlusions. Johannsen et al.[78] proposed a sparse decomposition of the epipolar plane image, which is performed over a depth-based dictionary built from fixed disparities, and deduces the scene disparity from the sparse coding coefficients. Jeon et al. [65] also used epipolar lines to correct distortions before applying stereo-matching techniques on the views of the light fields. Note that most of the aforementioned methods require an additional regularization or optimization step, which is usually a computationally intensive global process.

With the recent breakthrough of deep learning in computer vision research, applications of Convolutional Neural Network (CNN) to light fields have naturally arisen, including material recognition [79], view synthesis [80, 81] and super-resolution [82]. For depth map estimation from the light field, CNN is a popular method to address certain limitations of traditional methods, such as occlusions, specular highlights or reflections.

Heber et al. [83] first proposed to apply a conventional CNN in a sliding window fashion to predict the slop orientation in the epipolar plane images, which reduced the running time because of the property of CNN but still relies on an additional refinement step to handle textureless or uniform regions. A follow-up work [84] proposed an end-to-end u-shape network structure that operates on entire 2D epipolar plane images. This network achieved competitive results but suffered from streaking artefacts. To address these problems, an additional spatial regularization is introduced by [85] to perform 3D convolution operations on entire epipolar plane image volumes instead of independent epipolar plane images. However, this work still only consider the epipolar geometry from a single angular dimension of light field images when designing the network, resulting in low reliability of depth predictions. A multistream network is proposed by Shin et al. [86], which utilizes crosshair-shape geometric characteristics from each stack of images. Zhou et al. extend the structure crosshair stacks using multi-orientation epipolar plane image patches and adopt a multi-scale network for depth estimation [87]. Ma et al. introduce end-to-end network VommaNet for handling challenging reflective and texture-less areas when estimating accurate depth. This network features atrous convolutions of multiple scales and depth-wise convolution decreased parameter numbers [88]. Shi et al. propose to utilize several subsets of input views with different view selection strategies instead of structured input views to estimate depth from light fields. This method also demonstrates advantages on sparsely-sampled light fields with the large baseline [89]. Tsai et al. [90] employ the attention concept into the view selection strategy on the light field. In this way, views can be utilized more efficiently and accurately to perform depth estimation from the light field.

2.3 Light Field Reconstruction

Compared to sparsely-sampled light fields, densely-sampled light fields is generally advantageous for practical applications, including depth estimation, object segmentation and image-based rendering [91]. However, even with the blooming development of computational photography using modern digital devices, there still are some existing issues when we are capturing a light field using existing imaging systems. This issue is generally derived from the physical limitation of existing light field capturing systems. In our light field reconstruction research, we aim to develop algorithms to improve the limited quality of the collected light field data. In this section, we will introduce the formulation of light field reconstruction and summarize existing work on light field reconstruction.

2.3.1 Formulation of Light Field Reconstruction

To clarify the formulation of light field reconstruction, we will start from the analysis of light sampling in the ray space, which is derived from geometric analysis of light field rendering [92] and plenoptic sampling theory [93, 94, 95, 96]. Let's consider one continuous light field $\mathbf{L}^{continuous}(s, t, x, y)$, in which (s, t) denotes the camera plane and (x, y) denotes the image plane. The sampled light field $\mathbf{L}^{sampled}$ can be obtained by a sampling pattern function $s(\cdot)$:

$$\mathbf{L}^{sampled} = s(\mathbf{L}^{continuous}) \tag{2.10}$$

and corresponding reconstructed light field $\widehat{\mathbf{L}}^{reconstructed}$ can be obtained using a suitable reconstruction function $r(\cdot)$:

$$\widehat{\mathbf{L}}^{reconstructed} = r(\mathbf{L}^{sampled}) \tag{2.11}$$

The minimal sampling rate, which is also considered as the maximum camera spacing distance Δs_{max}, can be interpreted as the projection error from the reconstruction process [93]:

$$\Delta s_{max} = \frac{1}{2K_{f_x}} \tag{2.12}$$

where K_{f_x} represents the maximum frequency of the transformed frequency domain along the x dimension. To avoid overlapping between two adjacent sampling positions, Δs_{max} equals to Δx, i.e. one pixel, as if ignoring textural information to let $K_{f_x} = 1/(2\Delta x)$. In the other words, densely-sampled light fields are defined as having a disparity of less than *1px* between neighbouring views on the angular dimension [93]. This is intuitive as a pixel corresponds to the base unit size in the spatial dimension.

2.3.2 Light Field View Synthesis

To approach the ideal angular resolution of sparse light fields, one practical common solution is to increase the number of views via synthesizing novel views among sparse views. Note that in our work we adopt a looser definition of densely-sampled light fields, and rather focus on generating *denser* light fields from sparse inputs. In practice, this is limited by the existing hardware configuration of the acquisition systems mentioned in section 2.1.3. When capturing light fields, we are not capable of ensuring the quality of light field in both spatial and angular dimension.

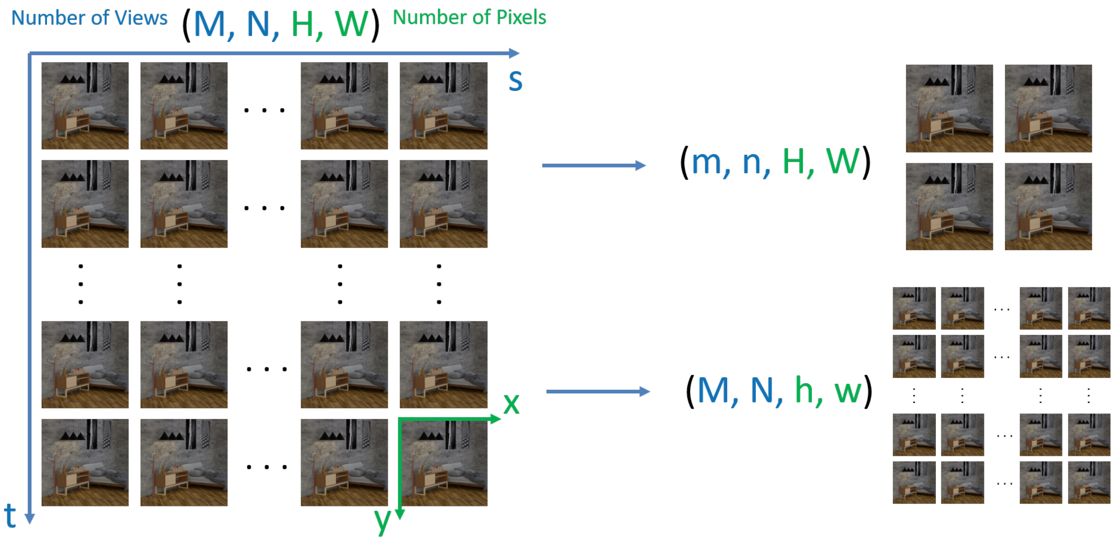


Figure 2.8: Trade-off of light field between number of pixels and number of views.

In the other words, the trade-off has to be made and the resolution has to be sacrificed on either spatial or angular domain. For example, camera arrays are usually built with dozens of high definition DSLR camera, which may provide ideal number of pixels (H, W) for each sub-aperture image but limited number of views (m, n). The number of DSLR cameras that can be set up limits the number of views as $M = \alpha(m-1)+1$ and $N = \alpha(n-1)+1$, in which α is the downscale factor in the angular domain. Moreover, the design of plenoptic camera based on a micro-lens array allows the user to capture light fields with high number of views (M, N), but only limited number of pixels (h, w), in which $H = \beta(h-1)+1$, $W = \beta(w-1)+1$ and β is the downscale factor in the spatial domain.

In this thesis, we will focus on reconstructing of sparse light field over the angular dimensions, typically captured with a camera array, via novel angular view synthesis. To formalize this problem, we can first define the dense 4D light field with high number of views as:

$$\mathbf{L}^D(s,t,x,y), (s,t) \in [1:M] \times [1:N], (x,y) \in [1:H] \times [1:W] \tag{2.13}$$

where $\mathbf{L}^D$ has a high number of pixels (H, W) and a high number of views (M, N). Similarly, a sparse 4D light field with low numeber of angular views can be depict as:

$$\mathbf{L}^S(s,t,x,y), (s,t) \in [1:m] \times [1:n], (x,y) \in [1:H] \times [1:W] \tag{2.14}$$

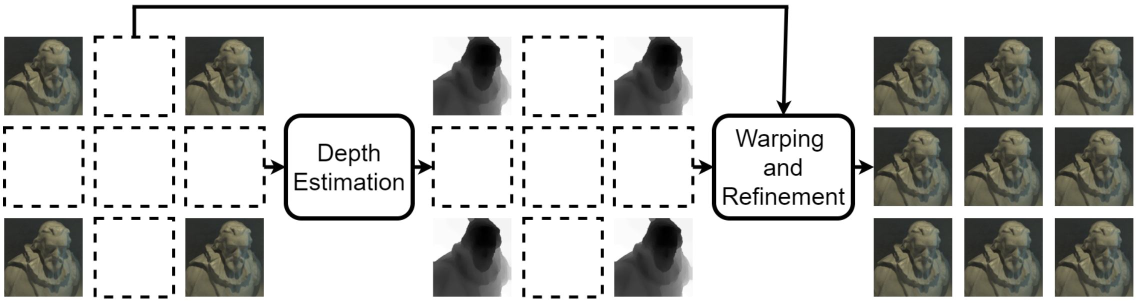


Figure 2.9: A generalized two-step depth based pipeline for light field view synthesis.

where $\mathbf{L}^S$ has low number of views (m, n) and equally well number of pixels (H, W).

The task of light field view synthesis, also called light field interpolation, is to establish an efficient and accurate framework, which is capable of super-resolving $\mathbf{L}^S(s, t, x, y)$ in the angular domain and obtain an angularly denser reconstructed light field $\widehat{\mathbf{L}}^D(s, t, x, y)$. The estimated dense light field is expected to be as close as possible to the original light field. Such a problem can be mathematically formulated as:

$$\widehat{\mathbf{L}}^D(s, t, x, y) = f_{VS}(\mathbf{L}^S(s, t, x, y)) \tag{2.15}$$

Where f_{VS} is the light field view synthesis framework. In the following sections of this chapter, We will introduce two different directions, depth-based and depth independent frameworks, to reconstruct densely-sampled light field from sparsely-sampled input.

Note that in previous work, light field view synthesis is sometimes associated with free-viewpoint rendering. In the rest of this thesis, unless specified otherwise, it will refer to the specific task of light field reconstruction or interpolation, *i.e.* when the synthesized views are positioned on the light field regular grid of viewpoints.

2.3.3 Depth Based View Synthesis for Light Field

Depth based view synthesis (or view interpolation) has been an active topic for many years, even before light fields were introduced. Conventional image-based rendering techniques [97, 98] consists of two steps: geometry estimation and view generation. When adapting this generalized framework to light fields, as shown in Figure 2.9, the first step is the estimation of the depth or disparity $\mathbf{D}$ from the light field, which is denoted as:

$$\mathbf{D} = f_d(\mathbf{L}^S) \tag{2.16}$$

Depth estimation from the light field is an active research area for which many existing approaches have been proposed, as described in section 2.2. The second step is to synthesize missing intermediate views using the estimated depth or disparity map:

$$\widehat{\mathbf{L}}^{D} = f_s(\mathbf{L}^{S}, \mathbf{D}) \tag{2.17}$$

A basic implementation of f_s consists in warping the closest input views to the target view and merge them, e.g. using bilinear weighted average.

A more advanced light field view synthesis was introduced by Wanner et al., who proposed a variational framework to generate novel views from sparse input views [99]. This work uses the structure tensor to perform disparity estimation, and then, input sub-aperture images are warped to synthesize novel views based on estimated disparity. An objective function is designed to be optimized to maximize the final generation quality. Although this optimization method is capable of producing plausible results in some light fields, various artefacts can still be observed on their synthesized results. The first reason is the loose connection between disparity estimation performed on the EPIs and the view synthesis performed on the sub-aperture images, which could cause information loss and reduce the quality of the final results subsequently. The second potential reason is Wanner et al.'s method assumes that light fields are acquired under perfect conditions. However, it is impractical to collect ideal images as most of the real-world images suffer from blur, noise and distortions.

The strength of deep learning based approaches has been demonstrated in many light field related methods, and the two steps of depth-based view synthesis can also be achieved by a combination of CNNs. Generally, one network would be responsible for estimating disparity information and the other is for reconstructing light fields. Kalantari et al. proposed the very first learning-based framework for light field view synthesis task, which aims to generate all intermediate sub-aperture views with only 4 input corner view [80]. This method consists of two cascade networks: the first CNN (f_d) estimates disparity on the location of each novel view from a set of features extracted from input views of the sparse light field, *i.e.* 4 corner views. The estimated disparity can be formulated as:

$$\begin{aligned} \mathbf{D} = f_d(&\mathbf{L}^{S}(1,1,x,y), \mathbf{L}^{S}(1,N,x,y), \mathbf{L}^{S}(M,1,x,y), \mathbf{L}^{S}(M,N,x,y)), \\ &(x,y) \in [1:H] \times [1:W] \end{aligned} \tag{2.18}$$

The second CNN (f_s) is then used as a color predictor to synthesize the target intermediate views using both the estimated disparity and the input views:

$$\widehat{\mathbf{L}}^{D}(s,t,x,y) = f_s(\mathbf{D}, \mathbf{L}^{S}(1,1,x,y), \mathbf{L}^{S}(1,M,x,y), \mathbf{L}^{S}(N,1,x,y), \mathbf{L}^{S}(M,N,x,y)),$$
$$(x,y) \in [1:H] \times [1:W] \tag{2.19}$$

These two CNN are connected and trained consecutively within each pass of the back-propagation with the ℓ_1-norm loss computed between synthesized views and ground truth views. Kalantari et al. used a set of 100 light field images captured from micro-lens array plenoptic cameras as the training dataset. The network is fed with equal-sized patches of 60×60 resolution. The final results of Kalantari et al.'s method are superior in comparison to [99] as the former method minimizes the error between the synthetic views and ground truth views directly, instead of optimizing depth map. However, it also can be observed that the results of Kalantari et al. [80] have failure cases when handling scenes under complex conditions such as occlusion, reflection, or large displacement.

Srinivasan et al. [100] provide a solution for light field view synthesis under a unique problem configuration, which aims to synthesize the sub-aperture view from a single RGB image with depth information detected by a depth sensor. This method synthesizes all views and corresponding depths at the same time, rather than processing each view at a time. Srinivasan et al. divide the problem of light field view synthesis to three tasks: the first task is to estimate the 4D ray depth at each location of views in the light field; the second task is to synthesize the lambertian approximation of the light field by utilizing the input single image and estimated depth from the previous task; the third task is to predict occluded rays and non-lambertian effects. Furthermore, these three tasks can be formulated as three different functions. The first one can be expressed as a function (f_s) in the equation (2.18). Given the estimated scene geometry from task one, the other two tasks can be formulated as a function f_{lam} and f_{occ}, respectively:

$$\mathbf{L}^{lam}(s,t,x,y) = f_{lam}(\mathbf{L}^{S}(\frac{1}{2}M, \frac{1}{2}N, x, y), \mathbf{D}),$$
$$(x,y) \in [1:H] \times [1:W] \tag{2.20}$$
$$\widehat{\mathbf{L}}^{D}(s,t,x,y) = f_{occ}(\mathbf{L}^{lam}(s,t,x,y), \mathbf{D}),$$
$$(s,t) \in [1:M] \times [1:N],\ (x,y) \in [1:H] \times [1:W] \tag{2.21}$$

Where $\mathbf{L}^{lam}$ represents the synthesized lambertian approximate light field, and $(\frac{1}{2}M, \frac{1}{2}N)$ represents the index of the single RGB input view at the center of the target light field. This work also published a dataset containing over 3300 light fields, which is captured using Lytro Illum camera and is claimed as the largest light field dataset available.

Recently, Jin et al. proposed an end-to-end learning approach that super-resolves sparse light field with large baselines [101]. This method contains three modules including a CNN based depth estimation module modelling the scene geometry, a backward-warping module synthesizing novel views at target locations, and another learning-based light field blending module fusing warped views for each location. Different from previous depth based view synthesis methods which perform fusion generally only using 2D convolution layers, the key feature of this method is employing the spatio-angular interleaved convolution [79, 102, 103, 104] into a light field blending module as the fusion step. That is, this blender is capable of modeling the angular coherency, which can be utilized to improve the reconstruction quality. Moreover, Jin et al. introduce a novel gradient loss computed on the epipolar-plane, which reveals the directional change of color intensity and promotes the geometry consistency of relevant light field parallax structure. This loss ℓ_e is defined as the ℓ_1-norm between the gradient of the synthesized epipolar plane image $\widehat{\mathbf{E}}$ and the ground-truth epipolar plane image $\mathbf{E}$:

$$\begin{aligned}\ell_e = \sum_{t,y}\Big(&\Big|\nabla_x \widehat{\mathbf{E}}_{t,y}(s,x) - \nabla_x \mathbf{E}_{t,y}(s,x)\Big| \\ &+\Big|\nabla_s \widehat{\mathbf{E}}_{t,y}(s,x) - \nabla_s \mathbf{E}_{t,y}(s,x)\Big|\Big) \\ +\sum_{s,x}\Big(&\Big|\nabla_y \widehat{\mathbf{E}}_{s,x}(t,y) - \nabla_y \mathbf{E}_{s,x}(t,y)\Big| \\ &+\Big|\nabla_t \widehat{\mathbf{E}}_{x,s}(y,t) - \nabla_t \mathbf{E}_{s,x}(t,y)\Big|\Big). \end{aligned} \tag{2.22}$$

This work focuses on super-resolving a sparse light field with 2×2 angular resolution to a dense light field with 7×7 angular resolution, and the evaluation demonstrates the advantages of this method when handling relatively wider baseline light fields.

As previous light field view synthesis methods mostly focus on the pixel level operation, Shi et al. [105] improved a general depth-based view synthesis framework by employing the feature level information for the reconstruction of light fields. A CNN is used to estimate the geometry of the scene, which is then used to warp input sparse views and their features to the target view locations. The key component of Shi et al.'s method is the fusion of reconstruction in two different spaces: pixel-wise and feature-wise space. Pixel wise reconstruction is similar to the refinement step from conventional depth based view synthesis pipeline, which is capable of modelling occlusions and in-painting on disoccluded areas via convolution layers. Feature wise reconstruction utilizes a coarse-to-fine VGG-19 network [106] to extract the perceptual information in multiple scales preserving high-frequency details. The whole network is trained in an end-to-end fashion so that extracted knowledge from these two modules can be shared, and the merge of these two domains brings an extra contribution to improve the quality of final reconstruction results.

Multi-plane image (MPI) is a new layered scene representation of the light field, which is originally derived from the plane sweep volumes (PSV) for the stereo magnification task [56]. The MPI representation can be used to generate novel views using homography to reproject each plane of the MPI to the desired viewport. Mildenhall et al. adapted this idea for light fields and proposed a practial solution to generate novel views from unstructured sparse light field, which is irregularly sampled and can be acquired via a smart phone [107]. The layered scene representation is promoted for each view of the light field using a 3D CNN, and novel views are then synthesized by combining intermediate generated views from the closest MPIs. Broxton et al. extend the concept of layered scene representation from planes to the spheres, which is introduced as the multi-sphere image (MSI) representation [38]. With specifically designed hardware configuration, MSI is capable of producing playback immersive video. However, the use of layered scene representation is computationally expensive, and the approach generates blurry results in the case of dynamic depth uncertainty.

All these aforementioned conventional depth-based light field view synthesis approaches utilize the estimated depth information to warp input views to generate and refine the novel views. On the contrary, learning-based view synthesis methods employ a neural network with supervised training to minimize the difference between the synthesized views and ground truth, instead of optimizing the depth map, giving rise to a higher quality of final results. However, all depth based methods still depend on the performance of their depth estimator, which perform best for light fields with small baselines and are prone to fail in challenging conditions, such as texture-less surface, non-lambertian surfaces or unstable illumination.

2.3.4 Depth Independent View Synthesis for Light Field

As there are various shortcomings of depth based methods, it is reasonable and crucial to explore depth independent methods for light field view synthesis. In order to enhance the angular resolution of light fields, various depth independent view synthesis methods are proposed and they can be utilized to generate dense intermediate views from sparse light field input. There are two main categories for the state-of-the-art depth independent methods: the first one is data-driven based methods; the other is domain transfer-based methods.

One common depth independent practice is to utilize the approximation ability of CNN, of which one example structure is shown in Figure 2.10. Features extracted by CNN may contain more comprehensive information than geometrical knowledge. Yoon et al. introduce the first depth-independent multi-stream neural network to improve both spatial and angular resolution [82]. The end-to-end structure and supervised training

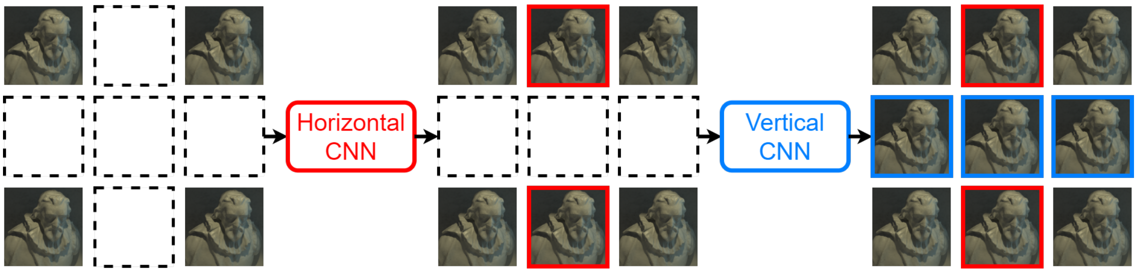


Figure 2.10: An example of the CNN structure for Light field view synthesis without depth estimation.

allow the parameters inside a network to be optimized by minimizing the difference between synthesized views and ground truth views, without considering to optimize a depth map. This method split input views to three different groups: horizontal, vertical and corners, which is fed to the corresponding input stream of the network. This design allows each stream of the network to follow the corresponding direction and also share the extracted geometrical features with each other. However, Yoon et al.'s network is only constructed with a limited number of plain convolutional layers, which leaves space for improvement, e.g. by employing advanced deep learning techniques, such as residual connection, batch normalization etc., which have been proven to greatly improve the performance of neural network on computer vision tasks.

Further research on deep learning-based light field view synthesis focus on exploiting the inherent consistency along the angular dimensions of the light field. Specifically, rather than extracting features using plain 2D convolution kernels on simply concatenated views, 3D or 4D convolution kernels are employed to enable high dimensional feature extraction. The significant advantage of the extra dimensions can also drastically increase the number of parameters and the global computational cost. Balanced strategies thus need to be designed when using such solutions, For this purpose, Wang et al. proposed an end-to-end network with pseudo-4D convolution to process 4D light field directly [108]. This pseudo convolution is implemented by efficiently combining a 2D convolution and a sequential 3D convolution. Given a input sparse light field $\mathbf{L}^S(s,t,x,y)$, this method perform the pseudo 4D convolution on a 3D volume $V_{t^*}(s,x,y)$, which consists of stacked sub-aperture views. This 3D volume is constructed by fixing one angular dimension of $t = t^*$. Yeung et al. introduced a two-step method, which first generates the whole set of novel views using a view synthesis network, and then retrieves intrinsic structure details using a view refinement network [103]. Different from Wang et al.'s method [108], Yeung et al. implement the network using an efficient spatio-angular alternating 2D convolution in a coarse-to-fine fashion, which can exploit the coherent structure on the 4D light field instead of the 3D volume. Inspired by the success of residual learning in image reconstruction area [109, 110], this method also employs the guided residual learning by connecting the input light field to various intermediate and final results.

The light field view synthesis problem could also be re-modelled as high-frequency details restoration problem in the epipolar plane image space. The first attempt is made by Wu et al. [111]. Given a input epipolar plane image $\mathbf{E}^S$ transformed directly from a sparse light field, Wu et al. consider the reconstruction of target reconstructed dense epipolar plane image $\widehat{\mathbf{E}}^D$ as formula below:

$$\widehat{\mathbf{E}}_H = f_{up}(\mathbf{E}_L) \tag{2.23}$$

Where f_{up} is the high-frequency up-sampling framework proposed by Wu el al. This framework consists of three cascaded modules: the first module is a blurring and up-sampling module, which is a Gaussian kernel k to extracts low-frequency information in the spatial domain. The next module is used to reconstruct details angular domain from blurred and up-sampled epipolar plane image exported from previous module. This module is implemented as a residual neural network f_{cnn} with three convolution layers, together with small-sized kernels as feature extractor and rectified linear units as the activation function. The last module D_k is designed to deblur and restore spatial details by employing a non-blind deblurring operation [112]. Thus, the proposed "blur-restoration-deblur" framework is assembled with these three modules, and is applied along both vertical and horizontal angular direction to fully reconstruct the dense epipolar plane images, and then build the complete 4D light field. The framework is optimized to solve the view synthesis problem by minimizing the following target function that describes the ℓ_2 distance between synthesized and ground-truth epipolar plane images:

$$\begin{aligned} & min \left|\left|\mathbf{E}_H - \widehat{\mathbf{E}}_H\right|\right|_2 \\ = & min \left|\left|\mathbf{E}_H - D_k(f_{cnn}(\mathbf{E}_L * k))\right|\right|_2 \end{aligned} \tag{2.24}$$

They also demonstrate the extension of their method on various applications, including depth enhancement, reconstruction of unstructured light fields and depth-assisted rendering [109]. It is worth to mention the depth-assisted rendering here, as it addresses the narrow baseline limitation of the original framework [111], which is caused by poor performance when using large blurring kernel k. The authors of [109] propose to discretize disparities and then apply appropriate shearing to corresponding epipolar plane image regions. In this way, large disparities can be transformed to a couple of small disparities, which satisfy the requirement of the original "blur-restoration-deblur" framework. Further, Wu et al. improved this work via replacing the depth assistance by a learning-based shearing module in [113]. This method disentangles the depth information by shearing the original epipolar plane images and convert the light field view synthesis problem to the fusion of differently-sheared epipolar plane images. In this way, the uncertainty of the depth estimation results can be avoided, and scene geometry information can be explicitly expressed by the properly sheared epipolar plane images. To select the suitable shearing value, an evaluation CNN is adapted in this work to estimated optimal sheared epipolar plane image structure. Wu et al. introduce a new similarity measurement between sheared epipolar plane image with ground truth epipolar plane image to find the optimal parameters of the evaluation CNN. Moreover, a coarse-to-fine fusion strategy is employed to integrate multiple reconstructed sheared epipolar plane image results eventually.

One common challenge of light field processing is collecting sufficient light fields that could approximate continuous properties of the target plenoptic function so that it can be reconstructed properly. One direction to resolve this issue is to find a suitable transform domain where dense light fields could become very sparse. Transform domain sparsification is a powerful prior, which can avoid the requirement for a huge amount of light field data. Shi et al. proposed an optimization framework in the continuous Fourier domain to reconstruct the dense light field [114]. The basis of this method is an important observation that the sparsity of light fields in the continuous Fourier spectrum is much greater than in the discrete Fourier spectrum, which is caused by the windowing effect. To formalize the problem, a signal $x(n)$ of length N is k-sparse in the continuous Fourier domain if it can be represented as a combination of $k, k < N$ continuous frequencies at arbitrary and non-integer locations:

$$x(n) = \frac{1}{N} \sum_{i=0}^{k} a_i exp(\frac{2\pi j t \omega_i}{N}), \tag{2.25}$$

where ${\omega_i}_k^{i=0}$ are the continuous positions of frequencies and ${a_i}_k^{i=0}$ are their corresponding coefficients. This formulation can be adapted to the 4D light field reconstruction problem, which is to recover the sparsity of the continuous spectrum from the 1D trajectory of the viewpoints. Given a sparse input light field $\mathbf{L}^S(s, t, x, y)$, a 2D slices $\mathbf{L}^S_{\omega_x,\omega_y}(s, t)$, which describes the fixed spatial frequency as a function of viewpoint, are obtained by applying 2D Discrete Fourier Transform on each input view individually. To reconstruct the dense light field, the 2D angular spectrum $\widehat{\mathbf{L}}^D_{\omega_x,\omega_y}(\omega_s, \omega_t)$ are recovered for each spatial frequency (ω_x, ω_y) from the input 2D slices. Later, a more advanced framework using the shearlet transform is proposed by Vagharshakyan et al. [115]. This method extended the concepts of light field sparsification and utilized the sparse representation of epipolar plane image in the shearlet domain, which exploit the anisotropic property of the epipolar plane image space. To handle the special characteristic of the epipolar plane image, the standard shearlet transform is modified to process straight lines. The results of this approach show better quality on challenging semi-transparent objects, compared to the results of [114]. Furthermore, Vagharshakyan et al. also develop a couple of accelerated extensions [116], which are based on the original method from [115]. Various techniques, including double overrelaxation, guided colorization and inter-EPI decorrelation, are utilized to enhance the final performance.

3 Disparity Estimation and Disparity based Light Field Reconstruction

In this chapter, we will introduce our contribution to light field disparity estimation and corresponding application to synthesize novel views for light field reconstruction. The background knowledge and related work, including optical flow estimation, edge-aware filtering and single image depth image-based rendering, are firstly outlined. More background details are described in section 2.2. Inspired by existing optical flow estimation and edge-aware filtering methods, we propose an accurate and efficient disparity estimation from light fields framework, producing geometrical information required by the consecutive view synthesis step. The optical flow estimator is applied on a sequence of images taken along an angular dimension of the light field, which produces several disparity map estimates. Considering both accuracy and efficiency, we choose the spatio-angular edge-aware method as our optical flow estimator. Thanks to its spatio-angular edge-aware filtering properties, the different disparity map estimates that we obtain are very consistent, which allows a simple one-step variational refinement to obtain the final disparity maps. Next, utilizing estimated disparity maps from the input light field, novel views are synthesized by warping input views to target locations. The evaluation is performed on both synthetic and real-world light field datasets and demonstrates the advantages of our disparity estimation and disparity based view synthesis method.

3.1 Related Work

3.1.1 Optical Flow Estimation

In computer vision, optical flow estimation is an active research domain. The original concept of optical flow is introduced to describe the motion trajectory of moving objects. Many methods have been proposed to estimate optical flow between frames considering both speed and accuracy, and each of them has to work under specific assumptions, which usually limits their performance. Horn and Schunck [117] introduced a pioneering energy minimization framework with a brightness consistency assumption, which became a common assumption for most of the following methods. Given this brightness consistency assumption, a generalized pixel relationship between two adjacent frames I_t, I_{t+1} from a continuous sequence could be formulated as:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{3.1}$$

and with correspondent Taylor series expansion and ignoring the high-order component, the equation becomes:

$$\frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt = 0 \tag{3.2}$$

where dx, dy, dt could be divided by dt at the same time, then the x and y components of the optical flow field at image $I(x, y, t)$ are obtained as $u = \frac{dx}{dt}$ and $v = \frac{dy}{dt}$:

$$\frac{\partial I}{\partial x} u + \frac{\partial I}{\partial y} v + \frac{\partial I}{\partial t} = 0 \tag{3.3}$$

which is known as Gradient Constraint Equation. However, mathematically, two unknown variables u, v can not be solved from one equation. Thus, additional conditions are required to solve this function. In [117], a global smoothness constraint is added with Gradient Constraint to establish an energy function:

$$E = \iint \left(\left(\frac{\partial I}{\partial x} u + \frac{\partial I}{\partial y} v + \frac{\partial I}{\partial t} \right)^2 + \lambda \left(|\nabla u|^2 + |\nabla v|^2 \right) \right) dxdy \tag{3.4}$$

which can be summarized to the classic form of energy function to be minimized for optical flow estimation:

$$E(J) = E_{data}(J) + \lambda E_{smooth}(J) \tag{3.5}$$

where J represents the motion vector (u, v) in the optical flow field. Directly solving such a large non-linear system is very challenging and usually time expensive. The class of global minimization problems requires optimizing a large number of independent variables, which usually leads to computationally expensive convergence. One direction to solve this problem is to replace variational optimization with the efficient edge-aware filtering process. Xiao et al. [118] integrated bilateral filtering into an iterative variational framework, replacing the traditional anisotropic diffusion step. More recently, iterative edge-aware filtering is introduced to efficiently approximate costly global regularization. Lang et al. [119] proposed an iterative spatio-temporal edge-aware filtering method which provides competitive accurate results and significantly reduced running time.

Some other methods aim to push the boundary of the accuracy aspect of optical flow estimation. With a focus on handling the local deformation details, Li et al. [120] propose a hybrid energy-based method which introduces a novel discrete Laplacian mesh energy concept as the additional term for core energy function expressed as:

$$E(J) = E_{data}(J) + \lambda E_{smooth}(J) + \xi E_{Lap}(J) \tag{3.6}$$

Another main issue that much optical flow work, including the original work from [117], suffer from is the inaccurate estimation when dealing with large pixel displacements. To improve the accuracy of such challenging cases, a patch-based similarity function is proposed [121] and integrated into data term, which was also initially introduced for the nearest neighbor field estimation. This patch match concept has been widely adapted to the optical flow problem. Patch match (PM) method usually could only provide sparse pixel correspondences, and the final optical flow estimation is then considered as a labelling problem leveraging the coherent information such as the geodesic distance of natural images. To further obtain a dense flow map, Revaud et al. [122] propose an edge-preserving interpolation scheme applied on the top of sparse matching correspondences. In this context, Hu et al. [123] propose a coarse-to-fine extension of the basic patch match method, which is proven efficient in finding reliable correspondences on large pixel displacements.

In addition to the accuracy in the spatial domain, the optical flow temporal consistency has been an important and challenging research topic. In the early work, Murray et al. [124], proposed to add a temporal smoothness term to improve the temporal consistency. Sliding windows [125] and Kalman filtering [126] based methods have been proposed later, focusing on temporal stability, however, performance of those approaches would highly rely on the selection of the window size. Feature flow (FF) [119] proposed a novel local edge-aware filtering to replace the expensive global optimization, which significantly reduced the computation cost while giving an accurate estimation. Temporally extended permeability filter (PF) [127] is proposed to produce competitive results with reducing halo artefacts efficiently. However, these filtering based methods usually rely on a sparse correspondence initialization, which has a significant impact on the final result.

3.1.2 Edge-aware Filtering

Edge-aware filters are important basic building blocks in many image and video processing methods, such as the aforementioned optical flow estimation. Barash et al. enable the multi-dimensional interpretation of the edge-preserving filters [128]. On a 2D RGB color image I, RGB values at pixel p with spatial coordinates (x_p, y_p) can be depicted as range coordinates $I(p) = (r_p, g_p, b_p)$. Consider $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$ as a 2D manifold M_I in $\mathbb{R}^5$, the quintuplet $\hat{p} = (x_p, y_p, r_p, g_p, b_p)$ represents a point on this manifold, which corresponds to pixel p on image I [129]. For each pixel p to be filtered, the generalized edge-aware filtering kernel $F(\hat{p}, \hat{p}_i)$ can be defined as:

$$J(p) = \int_{p_i \in \Omega} I(p_i) F(\hat{p}, \hat{p}_i) dp_i \tag{3.7}$$

where J is the filtered image, Ω is the window around pixel p and p_i is another pixel within the window range. In prior work, one remarkable edge-aware filter work is the bilateral filter, which is introduced by [130]. The bilateral filter replaces the intensity of each pixel with the weighted average of its neighboring pixels, which first provides the capability to preserve sharp edges while smoothing images. The bilateral filter kernel can be expressed by instantiating the generalized kernel $F(\hat{p}, \hat{p}_i)$ with the normalization term W:

$$F_{bilateral}(\hat{p}, \hat{p}_i) = \frac{1}{W} G_{\sigma_s}(||p - p_i||) G_{\sigma_r}(||I(p) - I(p_i)||) \tag{3.8}$$

where $||\cdot||$ is the measurement of ℓ_2 distance, G_{σ_s} and G_{σ_r} are Gaussian based spatial and range filters, with the support of coefficients σ_s and σ_r, respectively. Various edge-aware filters [131, 132, 133, 134, 135, 136, 137, 138] have been developed for different image-based applications such as stylization, HDR tone mapping and denoising. To address the drawbacks of the bilateral filter, the weighted least squares (WLS) filter [133] was shown to produce high quality filtering results and to suppress halo artefacts with a penalizing factor calculated from a distance between the original and filtered image. However, since it requires solving of a large linear system, the performance of this method is limited by its high computational cost. Later, edge-avoiding wavelets [134] were proposed to reduce the computational complexity, but suffer from aliasing problems and irregular edges because of the restricted size of kernels [138]. The local Laplacian filter is capable of producing high-quality results, but it also has the downside of being computationally demanding [137]. Aubry et al. provide a more efficient upgraded implementation of the Laplacian filter, but it is not clear how to extend the method into the temporal domain [138]. Besides, Criminisi et al. presented a geodesic-distance based framework for the luma channel in grayscale images [139]. More recently, the guided image filter exploits a local linearity assumption which considers the relationship between guidance I and filtered output J as a linear model [135], which can be formulated as:

$$J(p) = \frac{1}{N} \sum_{p_i \in \Omega} (a_p I(p_i) + b_p) \tag{3.9}$$

where J is the linear transform of I in a window region Ω centered at the pixel p. This model can be transformed into a similar form of generalized filter (3.7), and the guided filter kernel can be expressed as:

$$F_{guided}(\hat{p}, \hat{p}_i) = \frac{1}{N} \sum_{p_i \in \Omega} \left(1 + \frac{(I_p - \mu)(I_{p_i} - \mu)}{\sigma^2 + \epsilon} \right) \tag{3.10}$$

where N is the number of pixels in the window region Ω, and μ, σ are the mean and variance of all $I(p_i)$ in Ω. For the detailed mathematical derivation of this expression, we refer readers to read the supplementary material of [135]. Other state-of-the-art edge-aware filtering techniques such as the permeability filter [131, 140], the domain transform [132] and its extension for high-order recursive filtering [136] are all efficient techniques which offer a good quality-performance trade-off. However, as pointed out in [138], the guided image filter and domain transfer still suffer from halo artefacts, whereas the permeability filter does not, since it has specifically been designed to mimic similar behaviour as the high-quality weighted least squares filter, but with significantly lower computational complexity. Milanfar et al. [141] give an extensive overview of many filtering approaches.

3.2 Disparity Estimation and Disparity based View Synthesis on 4D Light Field

We propose a novel scheme for efficient and accurate estimation of disparity maps from the 4D structure of light fields using optical flow. Our approach consists of three main steps, as illustrated in Figure 3.1. First, input data is redefined as a 3D spatio-angular volume extracted from the light field by taking views along a given angular dimension. Second, the core algorithm of our framework is an optical flow estimation performed over this spatio-angular volume V. The displacements measured by the optical flow thus correspond to disparity estimates between consecutive views of the light field. The optical flow problem can be formulated as classic minimizing error energy function in

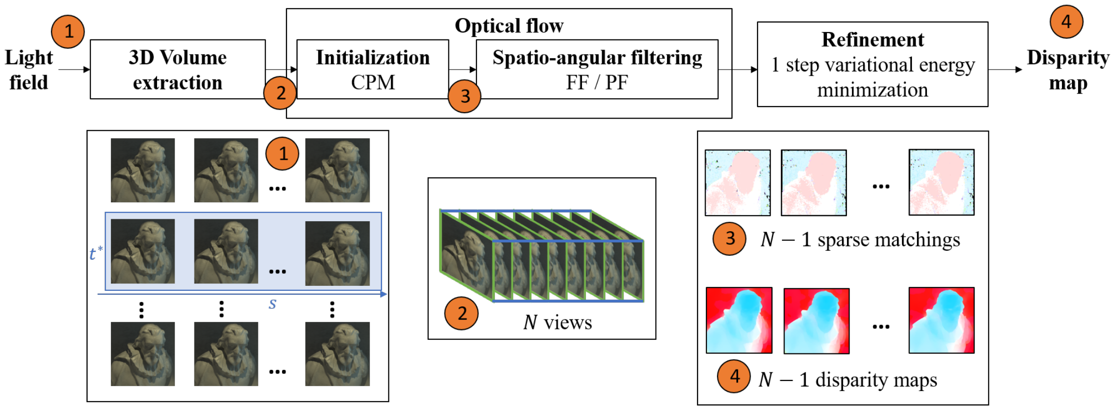


Figure 3.1: Overview of the proposed approach. The method can be applied on any number of rows or columns of the matrix of views in order to obtain more disparity map estimates.

the following form as below:

$$E(J) = E_{data}(J) + \lambda E_{smooth}(J)$$

where J represents the motion vector (u, v). Initialization with a sparse matching correspondence technique [123] is first performed to initialize $E_{data}(J)$ locally, then an efficient spatio-angular edge-aware filter replaces the expensive global optimization $E_{smooth}(J)$ to obtain a dense flow estimation. To accomplish this volumetric edge-aware filtering on the light field, we employed the feature flow method [119] and permeability flow method [127], which is suitable for multi-dimensional processing. Finally, a variational refinement step is performed to obtain the final disparity maps. Variational approaches are common in most state-of-the-art methods and perform computationally costly global energy minimization on the disparity map estimates to obtain the final accurate disparity map. Thanks to the edge-aware filtering along the angular dimension of the optical flow, we can reduce this step to a one-step variational energy minimization. Futhermore, we demonstrate the use of the estimated disparity maps in a light field view interpolation scenario.

3.2.1 3D Spatio-angular Volume Extraction

To obtain a 3D spatio-angular volume, we fix one of the angular dimensions $t = t^*$, as step 1 in Figure 3.1, and then extract N views over the remaining dimension s as a complete volume. This volume thus consists in a sequence of sub-aperture images along horizontal direction, noted $V = \{I_n\}, n = 1 \ldots N$. The same extraction operation can be performed along the vertical direction to obtain a corresponding disparity map. Here and for the rest of this chapter, we assume without loss of generality that we fix t^* and take the sub-aperture images over s.

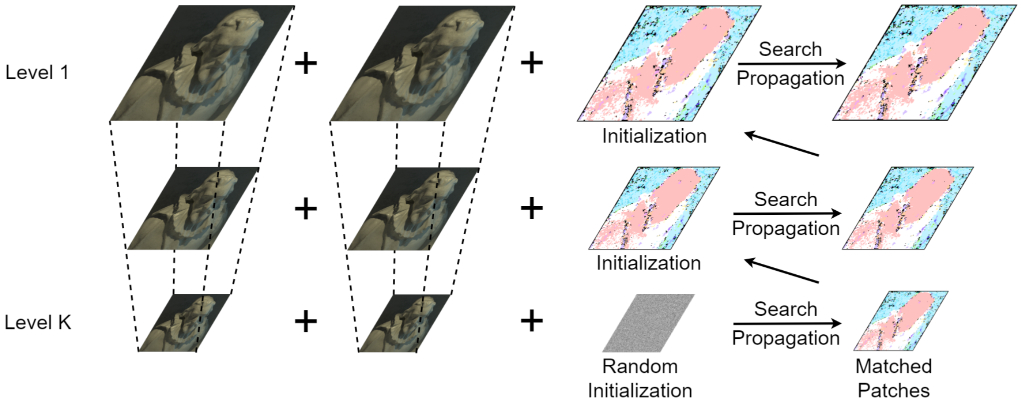


Figure 3.2: Coarse-to-fine Patch Match. Patches are initialized at the coarsest bottom level, and then patches matching are performed. Matched patches at each level will be used as the initialization for the next level.

3.2.2 Sparse Patches Matching Initialization

To calculate reliable quasi-dense sparse correspondences from views of the light field as initialization, we propose to use a recent extension of the PM method [121], the so-called Coarse-to-fine Patch Matching (CPM) technique [123], because it is both more efficient and accurate than the SIFT flow used in [119].

One level patch matching. We first introduce the basic matching algorithm on image pairs, which corresponds to two adjacent views from volume V. The CPM method aims to compute sparse matching between a pair of images, by applying a randomized nearest-neighbor field (NNF) algorithm on the pyramid constructed from the input image pair, as shown in Figure 3.2. The NNF method consists of three main components: initialization, propagation and search. First, the patches are paired with either random offset or inherited prior information from previous level. Considering the local smooth property of optical flow estimation, CPM performs matching only on selected seeds instead of on every pixel on input images. Given two views $\mathbf{I}_i, \mathbf{I}_j \in V$, we define a set of initial seeds $S = \{s_m\}$, and each seed is located at pixel location $p(s_m)$. Each seed corresponds to the central pixel of a regular square patch with a spacing of d pixels. The goal of this method is to determine the flow $f(s_m)$ at each location $p_i(s_m)$ in $\mathbf{I}_i$ pointing to a corresponding patch $\mathrm{M}(p_i(s_m)) = p_i(s_m) + f(s_m)$ in $\mathbf{I}_j$. Note that as we only look for sparse matches, the number of seeds is much lower than the total number of pixels. Second, coupled patches are iteratively examined in both scan order and reverse scan order. For a current seed s_m, the flow of its neighbor seeds will be propagated to s_m until the optimal flow is selected. This procedure can be formulated as:

$$f(s_m) = \underset{\mathrm{f}(s_a)}{\operatorname{argmin}}\ C(f(s_a), s_a), s_a \in s_m \cup N_m \tag{3.11}$$

where the cost function $C(\cdot)$ is designed to measure the difference between two patches from two images. In this method, it is implemented as the sum of absolute difference (SAD) of the SIFT descriptors, and N_m is a set comprising all patches examined in the current iteration. Last, to decide the final matching, a random search algorithm is applied for current s_m within a window region with radius r, in order to test some candidate flow around the current best flow. Note that in our context, the sub-aperture images of a light field are rectified, and we can further reduce the complexity of the matching search by limiting the search window to an epipolar line.

Pyramid Structure. The CPM method adapts the NNF method on a hierarchical architecture to estimate the optical flow. A pyramid with K levels is first constructed from the original images with a downsampling factor η. This pyramidal decomposition is noted $\mathbf{I}_i^k, \mathbf{I}_j^k$ with level $k = 1, \dots, K$. The seeds $\{s^k\}$ are also constructed on each level, we consider $\{s^k\}$ as the seeds on the k-th level, which is downscaled level by level, beginning from the first level:

$$s^k = \frac{1}{\eta} \cdot \{s^{k-1}\}, 2 \leq k \leq K \tag{3.12}$$

After the construction of image pyramids, we first perform the random initialization on the $\{s^K\}$, which is located on the bottom level. Then, the propagation and random search are performed to obtain the flow estimation $f(s^K)$ in the first level. Then, this obtained flow is utilized as the initialization for the next level. Such operation is iteratively performed on $\mathbf{I}_1^k$ and $\mathbf{I}_2^k$ with $k = K - 1 \dots 1$ using the output of the previous level $k + 1$ as initialization:

$$\{f(s^k)\} = \frac{1}{\eta} \cdot \{f(s^{k+1})\}, 1 \leq k \leq K - 1 \tag{3.13}$$

To initialize our optical flow, we apply the CPM method on consecutive pairs of views $\mathbf{I}_n, \mathbf{I}_{n+1}$ with $n = 1, \dots, N - 1$ taken from the volume V built previously, and we note F_n^{init} the flow between these views.

3.2.3 Edge-aware Filtering

Once sparse matches $\{F_n^{init}\}$ are obtained as described in the previous section, we perform edge-aware filtering on the estimated matches along the spatio-angular volume in order to obtain dense consistent correspondences. Here, we will describe two spatio-angular edge-aware filters: feature flow [119] and permeability filter [127], which can approximate and replace the global optimization process of traditional optical flow estimation.

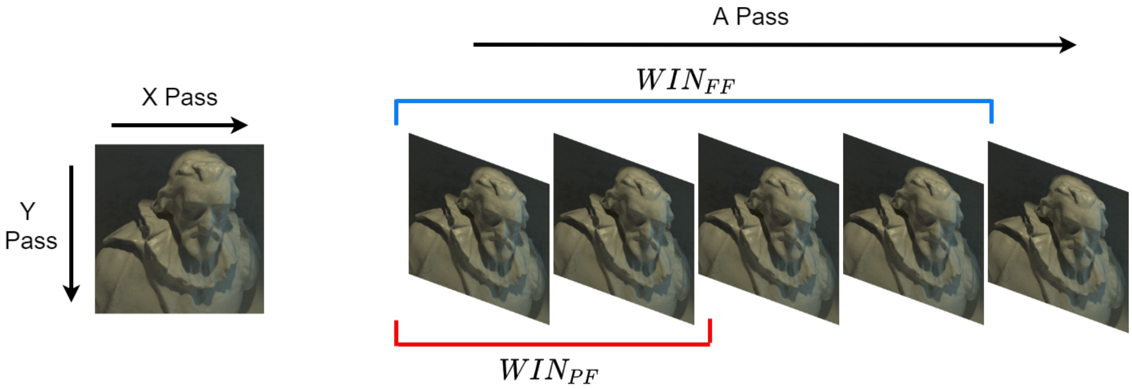


Figure 3.3: Spatio-angular edge-aware filtering. One iteration of the spatial filtering consists of one X Pass and one Y Pass. Angular filtering is performed within a sliding window along the angular dimension.

Spatio-angular Edge-aware Filtering #1: Feature Flow

Spatial Filtering. Inspired by the success of efficient edge-aware filtering process, we employ the feature flow method [119], an efficient edge-aware filter, to diffuse sparse matches with coherent information and obtained dense results. One of the main advantages of the feature flow is that the global energy minimization operation used in many optical flow approaches is replaced with a local volumetric edge-aware filtering operation. We first introduce how this method works on the spatial domain. To properly detect object edges in sub-aperture images and their disparity variations, a domain transform filter [132] is iteratively applied on the spatial dimension. Given this 2D image J, which is equal to the obtained matches $\{F_n^{init}\}$, $J : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defines the manifold M_J in $\mathbb{R}^4$. The spatial coordinates of the pixel p in J are denoted as (x_p, y_p), and the range coordinates of p is (u_p, v_p), as the target image is a flow image instead of a plain RGB image. Thus, this edge-preserving filtering process can be formulated as:

$$J_p^{XY(k+1)} = \sum_{q \in \Omega} J_q^{XY(k)} H(ct(p), ct(q)) \tag{3.14}$$

where J_p^k denotes the pixel range value at pixel p after k iterations of filtering, ct transforms p to the target domain, and H is a 1D kernel to filter a domain transferred 1D signal. The key idea of the domain transform filtering is to seek a suitable transform function ct, which is capable of transferring information to a lower-dimensional space and enables corresponding lower-dimensional filter kernel H while preserving high-frequency details in original domain. This construction is important as it could be more efficient when estimating ct and H than estimating an original high-dimensional kernel. According to the mathematical derivation from [132], by setting $ct(0) = 0$, the target domain transform ct can be expressed as:

$$ct(x) = \int_0^x 1 + \frac{\sigma_s}{\sigma_r} |J'(p)| \, dp, x \in J \tag{3.15}$$

where $J'(p)$ denotes the derivative of $J(p)$ with respect to p; σ_s and σ_r are the spatial and range parameters, which control the effect of the corresponding filters. When the spatial parameter σ_s increases, the kernel H tends to produce less smooth results. Furthermore, the amount of smoothing is inversely proportional to the gradient magnitude of the input. When the range parameter σ_r increases, the capability of retaining edge details decreases. Besides, to complete the filtering process on the 2D input matches, a succession of 1D filtering is performed, usually alternating between two passes scanning pixels along X-axes direction from left to right, and along Y-axes direction from top to bottom, respectively, until convergence and final processing of spatial filtered images $\{F_n^{XY}\}$, where $n = 1, \ldots, N - 1$.

Angular Filtering. After accomplishing the spatial filtering, the next step is to utilize coherent information along the angular dimension of the light field. To accomplish this goal, an angular filter inspired by the temporal filter from [119] is applied to the views from the volume V, which is extracted along one angular dimension. Given the spatially-filtered flows F_n^{XY} with $n = 1, \ldots, N - W + 1$, angularly-filtered are produced within a sliding window which contains W contiguous views.

To guide the filtering process, motion trajectories of pixels along the 1D angular dimension are used when applying the 1D filter kernel H. Starting from the first view $\mathbf{I}_1$, each pixel will start one path to be filtered within the sliding window W. However, different from the spatial domain where each row or column is well-defined, each motion path has to be strictly restricted to ensure its uniqueness for producing unbiased results. Thus, three rules are established as follows to determine the exclusive motion paths:

1. Once a path leaves the image boundary, the sliding window filter stops;

2. Once a pixel does not have previous paths mapped to it, a new path will be established from this pixel;

3. Once multiple paths collide at one pixel, we randomly allocate this pixel to one of those paths and let it continue, while terminating the others at the previous view;

One complete spatial-angular filtering process consists of one spatial filtering and one following angular filtering. For the first iteration of the angular filtering process, the sparse matching with one pass of spatial filtering F_n^{XY} is utilized as the initial pixel paths. After accomplishing each iteration of the angular filtering process, angular filtered results F_n^{XYT} will be used as the reference motion paths for the next iteration. Besides, as the scale of paths is variable, a double linked list is implemented and maintained to store paths with flexible length. Meanwhile, corresponding filter parameters σ_{ra} and σ_{sa} in the angular domain are also carefully selected.

Confidence normalization for the sparse-to-dense conversion. Although edge-aware filter is not designed as the interpolator, their mathematical properties allow them to spread sparse data, such as sparse matches $\{F_n^{init}\}$, following the guidance of image edges and producing dense results. This can be implemented first by involving a confidence map G as the normalization term. We initialize values of $G(x, y)$ as following principles:

$$G(x,y) = \begin{cases} 1, & \text{if } J(x,y) \text{ contains a valid value} \\ 0, & \text{otherwise} \end{cases} \tag{3.16}$$

The map G is subject to the same filtering operation and then is applied to the corresponding sparse data locations. The raw input $\{F_n^{init}\}$, which is to be filtered, is replaced by the dot product between G and $\{F_n^{init}\}$, which is depicted as: $G \cdot F^{init}$. After computing K iterations of filtering for both G and F, the map is utilized to normalize the filter output using a element-wise multiplication as $F^{XY} = (G \cdot F)^{XY(K)}./G^{(K)}$. Moreover, when estimating the optical flow or disparity map, this confidence term is capable of encouraging pixels with higher confidence score contribute more to the final result than low confidence pixels. To improve the accuracy of the flow estimation, we compute the confidence by measuring the difference between forward and backward matching, of which the value is normalized to the range 0.0 to 1.0. Additionally, we also integrate the occlusion handling function to the normalization map G by introducing a penalty function ρ:

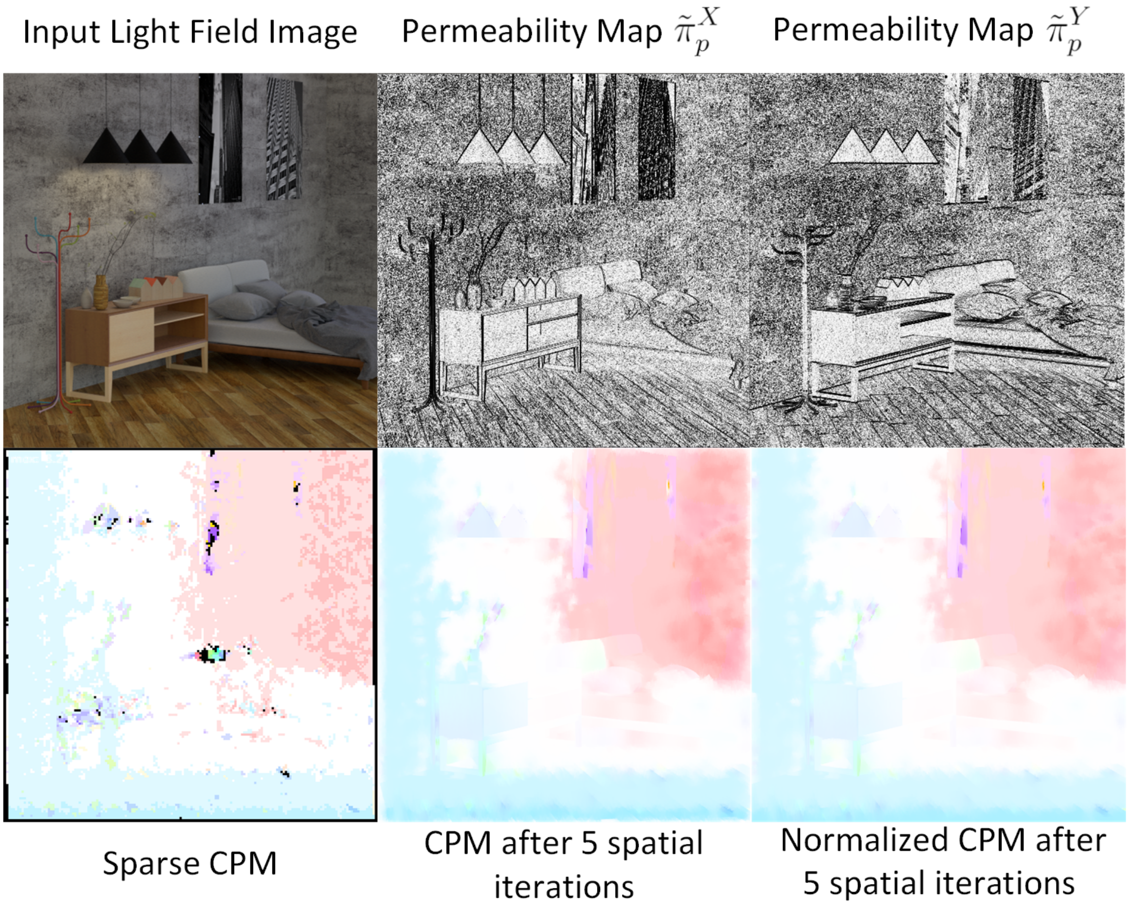
$$\rho = (1 - \left|\vec{w}^f + \vec{w}^b\right|)^\theta \tag{3.17}$$

where $\vec{w}^f$, $\vec{w}^b$ denotes forward and backward flow at each view, respectively,

and parameter θ determines the shape of the penalty curve. Beginning with the initial normalization G^0, the confidence penalty function is updated: $G^n = G^{n-1} \cdot \rho^{n-1}$ and applied to the input data: $J^n = J^{n-1} \cdot \rho^{n-1}$, before the filtering for each iteration. This term is capable of lowering the confidence of unreliable matching regions. In turn, this also increases the influence of those high confidence pixels on occlusion regions.

Spatio-angular Edge-aware Filtering #2: Permeability Filter

Although feature flow has already reached competitive performances in terms of accuracy and efficiency, some drawbacks remain when trying to achieve real-time applications on resource-limited devices. First, even for the optical flow estimation, an accurate optical



Input Light Field Image

Permeability Map $\tilde{\pi}_p^X$

Permeability Map $\tilde{\pi}_p^Y$

Sparse CPM

CPM after 5 spatial iterations

Normalized CPM after 5 spatial iterations

Figure 3.4: Spatial permeability filtering results on light field. Intermediate permeability maps and filtered results after 5 iterations are presented.

flow estimation is required at the beginning as prior alignment knowledge for angular filtering to initialize filtering iterations, which is difficult to obtain efficiently. Second, it has to operate iteratively along a complete image sequence or within a sliding window which leads to relatively high computation cost and extra memory use. Third, results from feature flow still suffer from typical halo artefacts, according to [138]. Therefore, one improvement direction of our previous disparity map estimation framework implementation is to replace feature flow method with a recently proposed *permeability filter*, which offers high quality and halo reduction with competitive speed. Besides, inspired by the mathematical derivation of permeability filter, the angular filtering process is also reformulated as an infinite impulse response filter, which can efficiently perform an incremental operation without costly motion path calculation or long-range frame alignment.

Spatial Filtering. The permeability filter is a type of edge-aware filter, which is usually used in a class of iterative problem of the following form:

$$J_p^{(k+1)} = \sum_{q \in \Omega} H_{pq} J_q^{(k)} + \lambda^{XY} H_{pp} (A_p - J_p^{(k)}) \tag{3.18}$$

where A_p denotes the unfiltered input image intensity at pixel position p at frame t, $J_p^{(k)}$ is the diffusion result at position p after k iterations. The set Ω contains all pixel positions of a frame, and H_{pq} are elements of the row stochastic matrix $H := \{H_{pq}\}$ defining the filter. Different from the aforementioned filters, the permeability filter adds one reference term A to guide the filtering process.

Given the obtained matches $\{F_n^{init}\}$, the iteration is initialized with $A = F_0^{init}$ and $J^0 = A$, where F_0^{init} represents the first frame of the sequence. Please note that all frame indices are omitted in this subsection since all operations are limited inside one single frame. The first term of equation (3.18) is the actual shift-variant convolution that denotes a diffusion estimate, and the second term is a fidelity term with $\lambda^{XY} \in [0, 1]$ which can be used to to induce bias in the iteration towards the input data A. We refer the reader to [131] for more details about the significant halo reduction induced by setting λ^{XY} to 1.

The permeability filter is a specific instance of equation (3.18) with H derived from two separate filter matrices H_X and H_Y for filtering operations in horizontal and vertical direction, respectively. These operations are applied independently, and a single spatial filtering iteration consists in a X pass followed by a Y pass. The two matrices H_X and H_Y are defined via permeability weights π_{pq} between two pixels p and q which control the local diffusion strength and show a low diffusion strength close to significant image edges. The permeability between two neighboring pixels $p = (x, y)$ and $p' = (x + 1, y)$ is defined as a variant of Lorentzian edge-stopping function:

$$\tilde{\pi}_p^X = \left(1 + \left| \frac{\|I_p - I_{p'}\|_2}{\sqrt{3} \cdot \sigma^{XY}} \right|^{\alpha^{XY}} \right)^{-1} \tag{3.19}$$

where I is the guiding image from original input data, σ^{XY} indicates the transition point from large to low permeability, and α^{XY} controls the transition rate of the edge-stopping function around σ^{XY}. In our implementation we use $\sigma^{XY} = 0.017$ and $\alpha^{XY} = 2$. Extending equation (3.19) to the general case, we can define the permeability between two arbitrary pixels p and q as :

$$\pi_{pq}^{X} = \begin{cases} 1 & \text{if } p = q \\ \prod_{n=p_x}^{q_x-1} \tilde{\pi}_{(n,p_y)}^{X} & \text{if } p_x < q_x, p_y = q_y \\ \prod_{n=q_x}^{p_x-1} \tilde{\pi}_{(n,p_y)}^{X} & \text{if } p_x > q_x, p_y = q_y \\ 0 & \text{else} \end{cases} \tag{3.20}$$

Then the final filter coefficients H_{pq} in equation 3.18 could be obtained by normalizing the pairwise permeabilities:

$$H_{pq} = \pi_{pq}^{X} \left(\sum_{n=1}^{w} \pi_{(n,p_y),q}^{X} \right)^{-1} \tag{3.21}$$

where w is the image width. Instead of following the normal steps that separately calculate the two terms in equation (3.18) and then sum them together, an efficient formulation for this filtering process is illustrated by [127, 140], which contains two-pass scan line operations, in both horizontal and vertical direction, with only constant computational complexity per pixel. In this formulation, the calculation of matrix H is actually avoided by using the permeability map instead, which only requires pairwise pixel calculation with lower computation complexity. After reformulating, one full filtering iteration consists of two steps: 1. left-right (top-bottom) pass; 2. right-left (bottom-top) and combination pass. Please note that the last combination operation can be efficiently carried out on-the-fly during the same loop of the right-left pass since all required variables are available after a right-left pass at each exact pixel position. Therefore, only two full scan passes are needed for the entire procedure and individual scan along one direction can be conveniently parallelized. In our current implementation, the initial values of all intermediate variables are all set to zero.

Besides, as edge-aware filters are not strictly interpolating filters, to efficiently perform a sparse-to-dense conversion, a normalization map G is also introduced. This map contains nonzero values at sparse sample positions and zero otherwise and is subject to the same filtering process applied to the sparse data channel. After the desired amount of iterations K, the element-wise normalization is performed between the filtered data and the filtered confidence map: $F^{XY} = F^{K}./G^{K}$. For the optical flow estimation, this normalization map is usually implemented as a feature matching confidence map which indicates the similarity between correspondences estimated by forwarding and backward flow. This confidence map increases the contribution of reliable correspondences and usually is normalized between 0.0 and 1.0. Please note that the original data has to be replaced by the product of the unfiltered data F and the initial confidence map G as $(F \cdot G)$ to meet the mathematical requirement before filtering. Visual results of the spatial permeability filter are shown in Figure 3.4. We can observe that after limited number of spatial iterations, the permeability filtering process can propagate sparse CPM to a dense disparity map and significantly removes outliers with a strong edge-aware diffusion.

Incremental Angular Filtering. Inspired by the iterative pairwise operation of the spatial permeability filter, the angular permeability filter is formulated as an infinite impulse response filter, improving both speed and memory use compared to previous filters [119, 131]. Given the spatial filtered results F_n^{XY}, angularly-filtered results F_n^{XYT} are produced within a sliding window which contains two contiguous views.

While the derivation of the angular permeability filter is built on an efficient formulation similar to the spatial permeability filter, there are two main differences to be noted. First, the inherent pairwise relationship between two adjacent pixels in the spatial dimension is replaced by pixel motion relationship, built by a forward mapping operation between pixels in two adjacent frames. Second, spatial permeability map is changed to a permeability combination of color consistency and a flow-gradient magnitude measured in the angular direction. The photo constancy is a straightforward extension of the spatial permeability:

$$\tilde{\pi}_t^{photo} = \left(1 + \left| \frac{\left\| \mathbf{I}_t - warp_{\mathbf{F}_{t-1}^{XYT}}(\mathbf{I}_{t-1}) \right\|_2}{\sqrt{3} \cdot \sigma^{photo}} \right|^{\alpha^{photo}} \right)^{-1} \tag{3.22}$$

which allows angular filtering along motion paths with similar color values between two adjacent images. To prevent angular filtering from stopping at warping errors and artefacts resulting from angular color inconsistency, the gradient-magnitude measure is calculated as:

$$\tilde{\pi}_t^{grad} = \left(1 + \left| \frac{\left\| \mathbf{F}_t^{XY} - warp_{\mathbf{F}_{t-1}^{XYT}}(\mathbf{F}_{t-1}^{XYT}) \right\|_2}{\sqrt{2} \cdot \sigma^{grad}} \right|^{\alpha^{grad}} \right)^{-1} \tag{3.23}$$

where α and σ are control parameters similar to the spatial filter parameters. Divisions and exponentiations are all element-wise in the two equations above and the final angular permeability is obtained by multiplying the two terms: $\tilde{\pi}_t^T = \tilde{\pi}_t^{photo} \cdot \tilde{\pi}_t^{grad}$. In our implementation, targeting optical flow estimation, parameters values are set to $\sigma^{photo} = 0.3$, $\sigma^{grad} = 1.0$, $\alpha^{photo,grad} = 2$. Visual results of angular permeability filter are shown in Figure 3.5. As we can observe, gradient measure can reveal regions where warping artefacts are likely to happen.

SABOM descriptor. Additionally, when employing permeability filter, we also improve

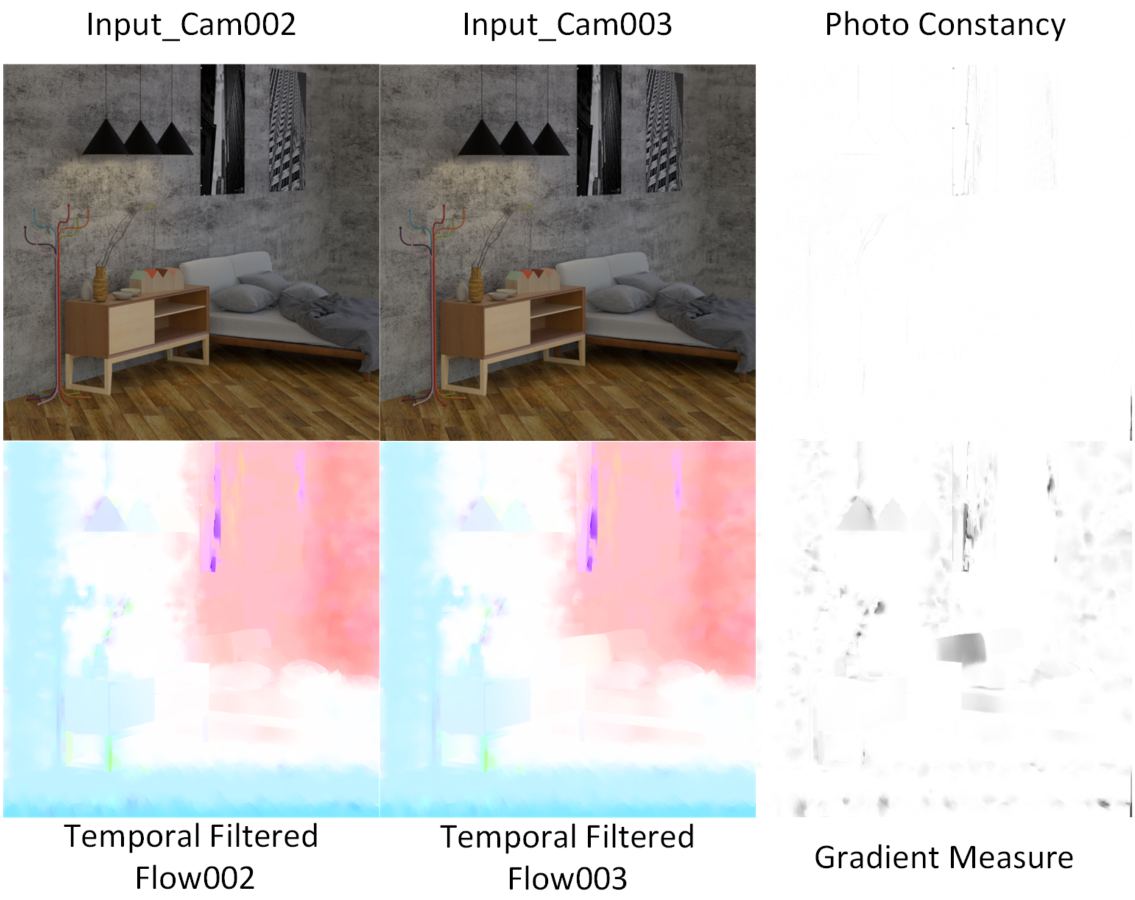


Figure 3.5: Angular permeability filtering results on light field. Photo constancy and gradient measures for the light field are presented.

the matching process by replacing the SIFT descriptor with the binary SABOM descriptor recently proposed by Alain et al. [142]. This spatio-angular binarised orientation maps (SABOM) descriptor extends the binarized octal orientation maps (BOOM) descriptor proposed by [127] by exploiting the light field gradient over both the spatial and the angular dimensions, which was shown to significant reduction of processing time while keeping competitive accuracy performance compared to SIFT [142].

3.2.4 Variational refinement

Thanks to the angular filtering which enforces the consistency between the different disparity map estimates, we can apply a simple one-step variational energy minimization, as opposed to costly global energy minimization technique, on the spatio-angularly filtered results F_n^{XYT}. We used the successive over relaxation method [143], also adopted by the Epicflow [122], in order to obtain the final disparity maps $\mathbf{D}$:

$$\mathbf{D} = \arg\min_{F_i^{XYT}} (E_{data}(F_i^{XYT}) + \alpha E_{smooth}(F_i^{XYT})) \tag{3.24}$$

where E_{data} corresponds to a classical color-constancy data term while E_{smooth} corresponds to a gradient-constancy function with a local smoothness term weight $\alpha = exp(-\kappa \|\nabla_2 \mathbf{D}\|)$ [144], where $\kappa = 5$.

3.2.5 Disparity based View Synthesis

The disparity maps estimated through our approach can further be used for view synthesis purpose. We give here a description of the view synthesis method employed to generate our results, which was briefly mentioned in section 2.3.3, and consists in warping the closest input views to the target view, and merge them using a bilinear weighted average. Note that we are able to use this view synthesis approach thanks to the fact that our disparity estimation method provides a disparity map for each input view of the light field, which is not the case for most disparity or depth estimation methods, which often only estimate the geometry for the centre view.

Formally, we denote by $\mathbf{I}_u$ the unknown target view to be synthesized at position (s_u, t_u), $\mathbf{I}_k, k = 0 \ldots 3$ the 4 closest known views in the input light field at positions (s_k, t_k), and $\mathbf{D}_k$ the corresponding disparity maps obtained from our approach. Note that the target view position (s_u, t_u) can be arbitrary, however in our results we use this view synthesis approach in a light field interpolation application where the target view positions fall onto the light field grid of views.

First, each known input view $\mathbf{I}_k$ is warped to the target view position using a warping operator $\mathbf{W}$:

$$\widehat{\mathbf{I}}_u^k = \mathbf{W}_{s_k,t_k \longrightarrow s_u,t_u}(\mathbf{I}_k, \{U_k, V_k\}) \tag{3.25}$$

where $\{U_k, V_k\}$ is a 2D warping flow denoted as:

$$\{U_k, V_k\} = \{(s_u - s_k) \times \mathbf{D}_k, (t_u - t_k) \times \mathbf{D}_k\} \tag{3.26}$$

The final estimate is then obtained as a weighted average of the 4 initial warped estimates:

$$\widehat{\mathbf{I}}_u = \frac{1}{\sum_{k=0,3} w_k} \sum_{k=0,3} w_k \widehat{\mathbf{I}}_u^k \tag{3.27}$$

where the weight are bilinear coefficients such that $w_k = (1 - |s_k - s_u|) \times (1 - |t_k - t_u|)$.

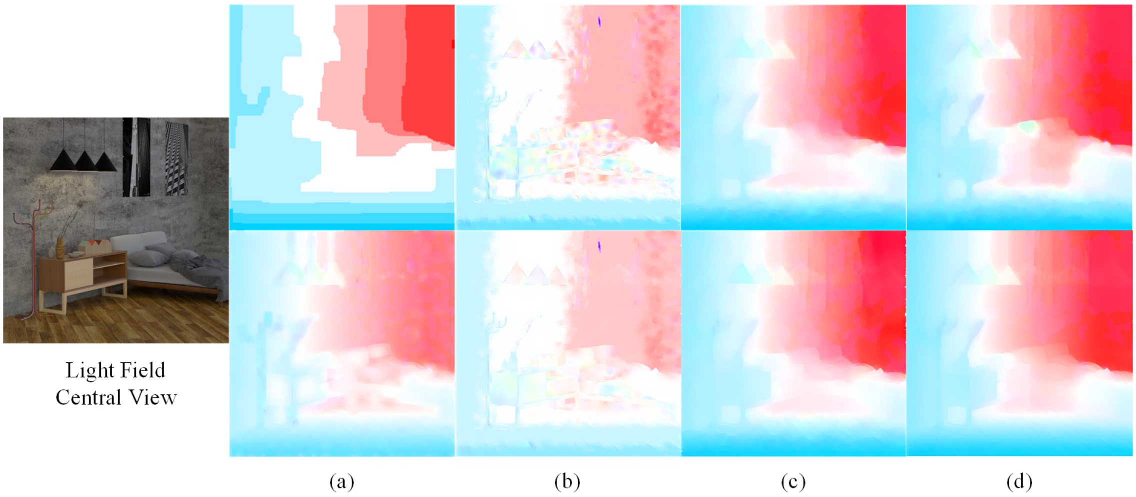


Figure 3.6: Comparison of optical flow with state-of-the-art methods. **Top row** consists of initialization results with different optical flow methods. **Bottom row** is the results of these initializations + feature flow filter. (a) SIFT Flow [145] (4.9s); (b) EPPM [146] (GPU-based,0.7s); (c) EpicFlow [122] (15s); (d) CPM-Flow [123] (5.3s)

3.3 Results and Evaluation

In this section, we analyze the results of the proposed approach. All our experiments were run on an Intel Core i7-6700k 4.0GHz CPU. We use the feature flow implementation from [147] and implement our own version of permeability filter. All parameter settings are retained from the original papers and maintained across all our experiments. Note that CPM is implemented with SIFT descriptor for feature flow (CPM_FF) and with SABOM descriptor for permeability filter (CPM_PF).

3.3.1 Evaluation of Disparity Estimation from Light Field

Evaluation of the optical flow initialization.

We evaluate here the performance of the proposed optical flow approach against state-of-the-art methods. In Figure 3.6, we show the results of several optical flow initializations in the top row and the results after feature flow filtering in the bottom row. The volumetric filtering using feature flow along the angular dimension of the light field clearly improves the accuracy of the optical flow from any initialization method, significantly improving consistency and continuity of brightness. The proposed method using CPM as initialization achieves the best performance in terms of the balance between speed and accuracy. The importance of the volumetric filtering is also illustrated in Figure 3.8, where we show the final depth map results for several light fields obtained with our method with and without feature flow. The quality of the depth maps is clearly improved for all sequences.

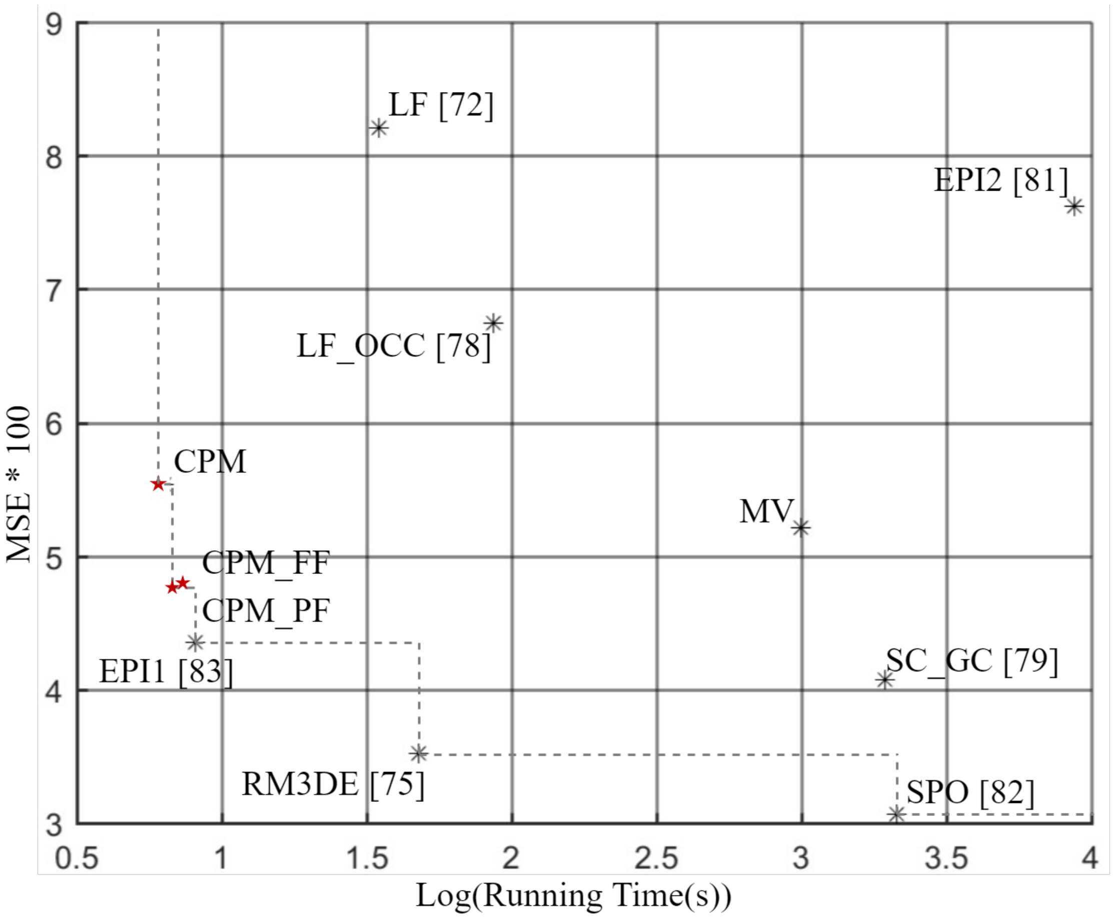


Figure 3.7: Comparison of our methods (red stars) performances against state-of-the-arts (blue stars), averaged over all HCI light fields. The results show that we achieve comparable performances to the best state-of-the-art method in terms of the balance between speed and accuracy.

Evaluation of the depth estimation on HCI dataset.

We evaluate here the accuracy and efficiency of our proposed method against state-of-the-art light field depth map estimation methods [65, 69, 72, 73, 76, 77, 78] using the HCI 4D light field dataset [3] [1]. Note that since this benchmark expects a depth map and not a disparity map for its evaluation, the disparity maps **D** obtained from our approach are converted to depth maps using equation 2.9. The accuracy of the depth estimation is evaluated using the Mean Square Error (MSE) * 100 and the computational complexity using the running time in seconds. The results are summed up in the graph of Figure 3.7, showing the average performances over the HCI dataset. Our method achieves comparable performance with the best method of the state-of-the-art in terms of the balance between accuracy and speed.

In addition to these objective metrics, we show the depth maps obtained from several

[1] http://hci-lightfield.iwr.uni-heidelberg.de/

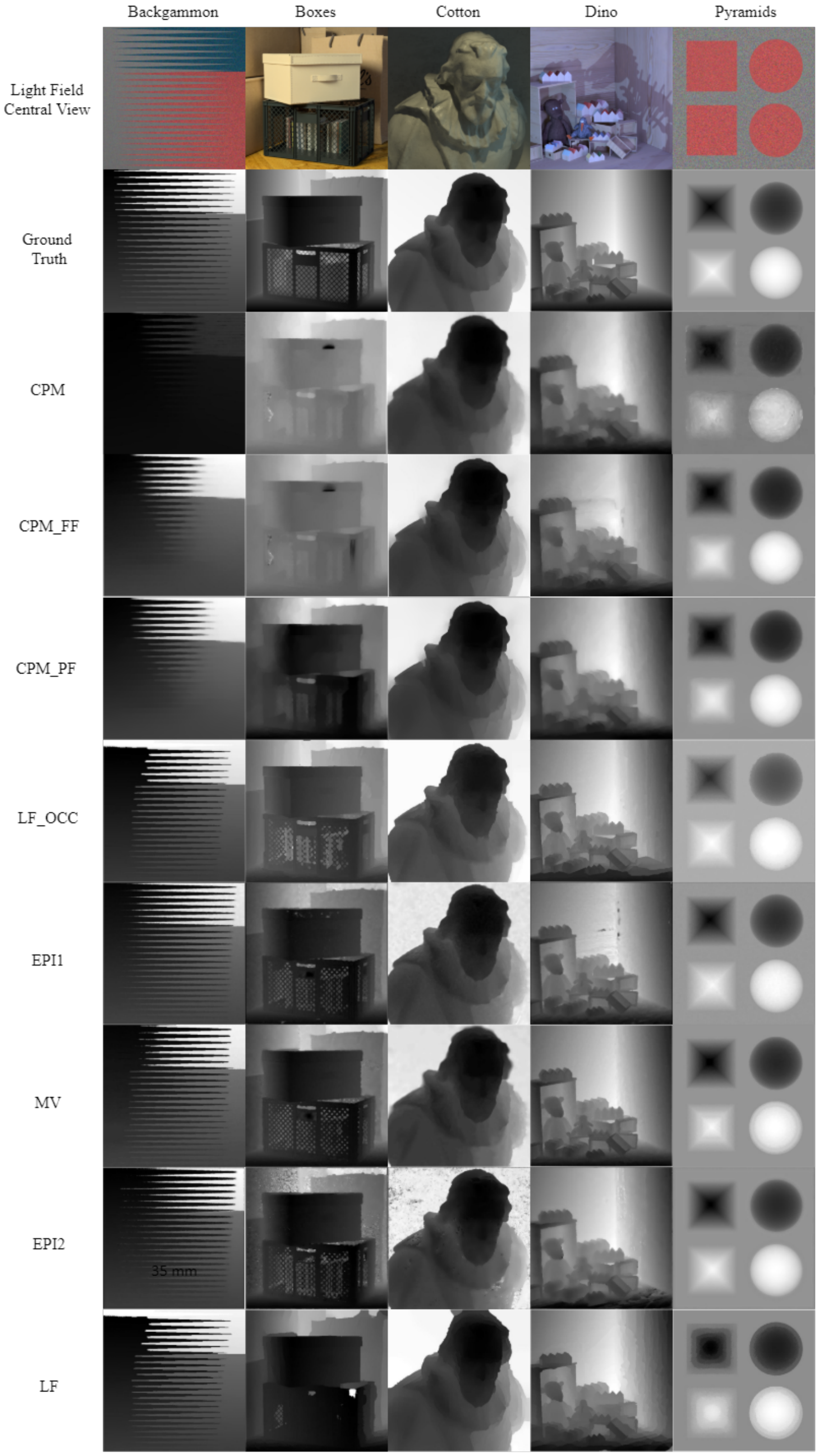


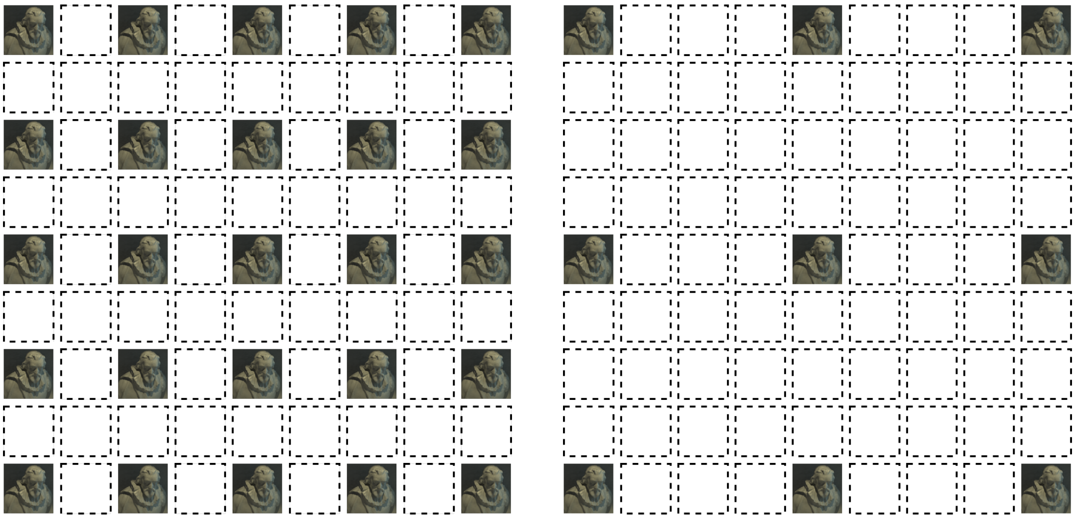
Figure 3.8: Depth map comparison on HCI dataset.

light fields in Figure 3.8 and compare against state-of-the-art methods. The final comparison shows better performance for edge preservation of objects (see Cotton column) and also smoother results for noisy scenes (see Backgammon and Dino columns). However, we notice that the proposed local filtering method, which allows considerable speed up, sometimes produces less smooth results than a global solution (see the background of Dino and Boxes column). The feature flow filter also heavily depends on the quality of the optical flow initialization. If the optical flow method is unable to provide accurate correspondences, it can not be corrected by the filter (see for example the Boxes column).

3.3.2 Evaluation of Disparity-based View Synthesis for Light Field

We assess here the use of our disparity map estimation in a light field interpolation scenario, using the disparity-based view synthesis (DVS) method described in section 3.2.5. For this purpose, we sub-sample existing light fields with a factor $\alpha = 2$ or 4, in order to have a ground truth reference for the interpolated views, as shown in Figure 3.9.

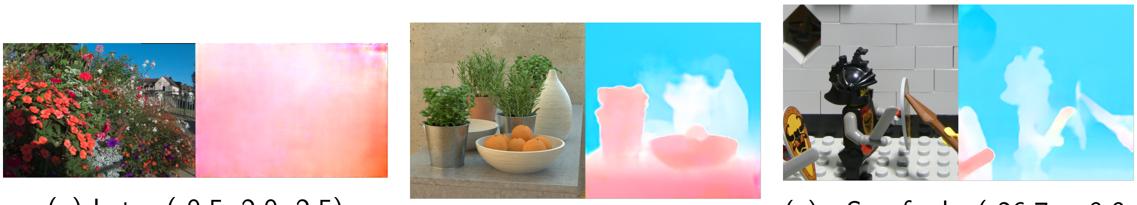
Here we evaluate the disparity-based light field view synthesis against a baseline angular bilinear interpolation, however note that more comparisons against state-of-the-art view synthesis and light field view synthesis methods are provided in the next chapters. The numerical results on Lytro, HCI and Stanford datasets are presented in the Table 3.1, 3.2 and 3.3, respectively. We use the objective metrics PSNR and SSIM in our evaluation to measure pixel and structural quality. Note that we report the metric values averaged over all interpolated views.



(a) $\alpha = 2$

(b) $\alpha = 4$

Figure 3.9: Two interpolation scales of the disparity based view synthesis.



(a) Lytro (-0.5, 2.0, 2.5)

(b) HCI (-13.7, 7.6, 21.3)

(c) Stanford (-26.7, 9.9, 36.6)

Figure 3.10: Disparity range (d_{min}, d_{max}, d_{range}) of Lytro, HCI and Stanford datasets ($\alpha = 4$).

We evaluate the disparity range of three used light field datasets, in which d_{min} and d_{max} stands for the minimal and maximum disparity (pixel) over all input views and disparity range (d_{range}) stands for the range of disparity (pixel) between d_{min} and d_{max}. Figure 3.10 illustrates the disparity range among each of the sampled test light fields. Even with larger sampling factor ($\alpha = 4$), the disparity of the test light fields from the Lytro dataset is only ~ 2 pixels, which is much smaller than test light fields from the HCI dataset (~ 21 pixels) and the Stanford dataset (~ 36 pixels). With such limited disparity, the bilinear interpolation is sufficient to obtain numerically good results on the Lytro dataset.

Meanwhile, depth based method may cause error due to potential mismatched correspondences, which makes results worse than the results from bilinear interpolation without disparity information involved on the Lytro dataset. However, the advantage of the disparity-based view synthesis appeared on the sparser HCI and Stanford datasets, with a $\sim$ 6dB gain in PSNR in average for the HCI dataset and $\sim$ 3dB when $\alpha = 2$ to $\sim$ 4dB gain in PSNR $\alpha = 4$ in average for the Stanford dataset. The visual results for $\alpha = 4$ are presented in Figure 3.11 and confirm the numerical results. We selected four representative light fields: *Herbs* and *Bicycle* are from the HCI dataset containing complex indoor scene, while *ChezEdgar* is from Lytro with small disparity, and *LegoKnights* is the most challenging image from Stanford with larger disparity and textureless surfaces.

While little difference can be observed between the ground truth, bilinear, and disparity-based view synthesis on the Lytro *ChezEdgar* light field, the disparity-based view synthesis provides obviously better visual results than the bilinear interpolation for the HCI and Stanford datasets. However, artifacts can still be observed in the disparity-based view synthesis results, especially on more challenging scenes such as the Stanford *LegoKnights* light field, which motivates our following work presented in the next chapters.

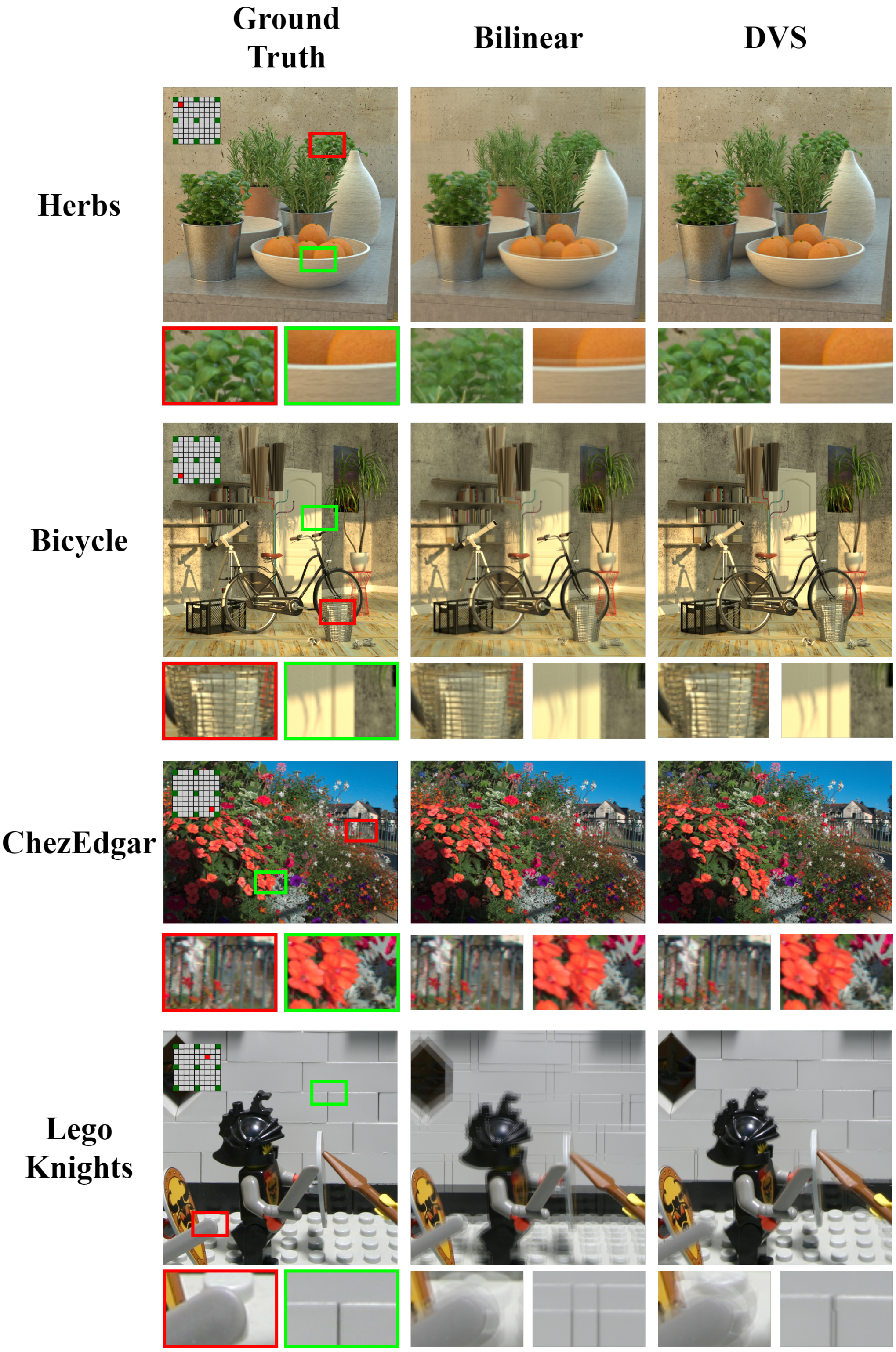


Figure 3.11: Comparison of our results against state-of-the-arts on HCI dataset. The disparity based light field view synthesis uses our estimated disparity with CPM_PF method.

Table 3.1: Numerical results on the real-world Lytro datasets [1, 2]

$\alpha = 2$	PSNR(dB)					SSIM				
	average	Bee_2	Bikes	ChezEdgar	Desktop	average	Bee_2	Bikes	ChezEdgar	Desktop
Bilinear	**37.33**	**35.49**	**37.98**	**38.03**	**37.82**	**0.9870**	**0.9705**	**0.9932**	**0.9925**	**0.9918**
DVS	36.40	35.31	37.60	37.44	35.25	0.9866	0.9702	0.9930	0.9920	0.9912

$\alpha = 4$	PSNR(dB)					SSIM				
	average	Bee_2	Bikes	ChezEdgar	Desktop	average	Bee_2	Bikes	ChezEdgar	Desktop
Bilinear	**32.60**	**30.29**	32.50	**34.12**	**33.10**	**0.9753**	**0.9010**	0.9721	0.9835	0.9758
DVS	32.25	30.18	**32.53**	34.02	32.27	0.9594	0.8998	**0.9745**	**0.9842**	**0.9793**

Table 3.2: Numerical results on the synthetic HCI dataset [3]

$\alpha = 2$	PSNR(dB)					SSIM				
	average	bedroom	bicycle	herbs	origami	average	bedroom	bicycle	herbs	origami
Bilinear	31.45	33.37	30.71	28.76	32.97	0.9397	0.9672	0.9649	0.8586	0.9682
DVS	**37.67**	**41.08**	**34.80**	**35.82**	**38.98**	**0.9931**	**0.9945**	**0.9929**	**0.9880**	**0.9972**

$\alpha = 4$	PSNR(dB)					SSIM				
	average	bedroom	bicycle	herbs	origami	average	bedroom	bicycle	herbs	origami
Bilinear	28.04	30.09	26.82	26.69	28.55	0.8733	0.9069	0.8909	0.7760	0.9194
DVS	**34.94**	**39.44**	**31.81**	**33.15**	**35.37**	**0.9855**	**0.9931**	**0.9824**	**0.9732**	**0.9933**

Table 3.3: Numerical results on the real-world Stanford Gantry datasets [4]

$\alpha = 2$	PSNR(dB)			SSIM		
	average	LegoKnights	TheStanfordBunny	average	LegoKnights	TheStanfordBunny
Bilinear	32.57	26.17	38.98	0.9141	0.8463	0.9819
DVS	**35.67**	**30.64**	**40.70**	**0.9765**	**0.9642**	**0.9888**

$\alpha = 4$	PSNR(dB)			SSIM		
	average	LegoKnights	TheStanfordBunny	average	LegoKnights	TheStanfordBunny
Bilinear	28.60	22.85	34.35	0.8441	0.7398	0.9484
DVS	**32.87**	**27.51**	**38.23**	**0.9586**	**0.9303**	**0.9869**

3.4 Conclusion

In this chapter, we introduced a novel optical flow-based method to estimate depth maps from light fields. We showed that by extracting a 3D volume consisting of a sequence of views from the 4D light field and applying a temporally consistent optical flow on this spatio-angular volume, we were able to obtain high-quality disparity maps and subsequent rendered views with reduced complexity. Comparison with state-of-the-art depth estimation methods on the HCI benchmark showed that we are competitive with the best methods in terms of the balance between accuracy and speed. The results of disparity based view synthesis also demonstrate the success except for the large baseline light fields.

4 A Study of Efficient Light Field Subsampling and Reconstruction Strategies

In this chapter, we investigate subsampling and reconstruction strategies for light fields. The two plane representation of light fields is adopted in this work to perform subsampling and reconstruction, as shown in Figure 4.1. Limited angular resolution is one of the main obstacles for practical applications of light fields. Although numerous approaches have been proposed to enhance angular resolution, view selection strategies have not been well explored in this area. More background details are described in section 2.3. This chapter looks at the view selection strategies from a subsampling and reconstruction viewpoint. First, different subsampling strategies are studied with a fixed sampling ratio, such as row-wise sampling, column-wise sampling, or their combinations. Second, several strategies are explored to reconstruct intermediate views from four regularly sampled input views. The influence of the angular density of the input is also evaluated. We evaluate these strategies on both real-world and synthetic datasets, and optimal selection strategies are devised from our results. These can be applied in future light field research such as compression, angular super-resolution, and design of camera systems.

4.1 Related Work

4.1.1 View Selection Strategies for Light Field View Synthesis

Having multiple sub-aperture images that are captured with similar angles, light fields offer some advantages compared to traditional single images. To find the optimal view selection strategy for practical applications, we should make use of the inherent similarity among multiple angular views. However, most previous methods only focus on one

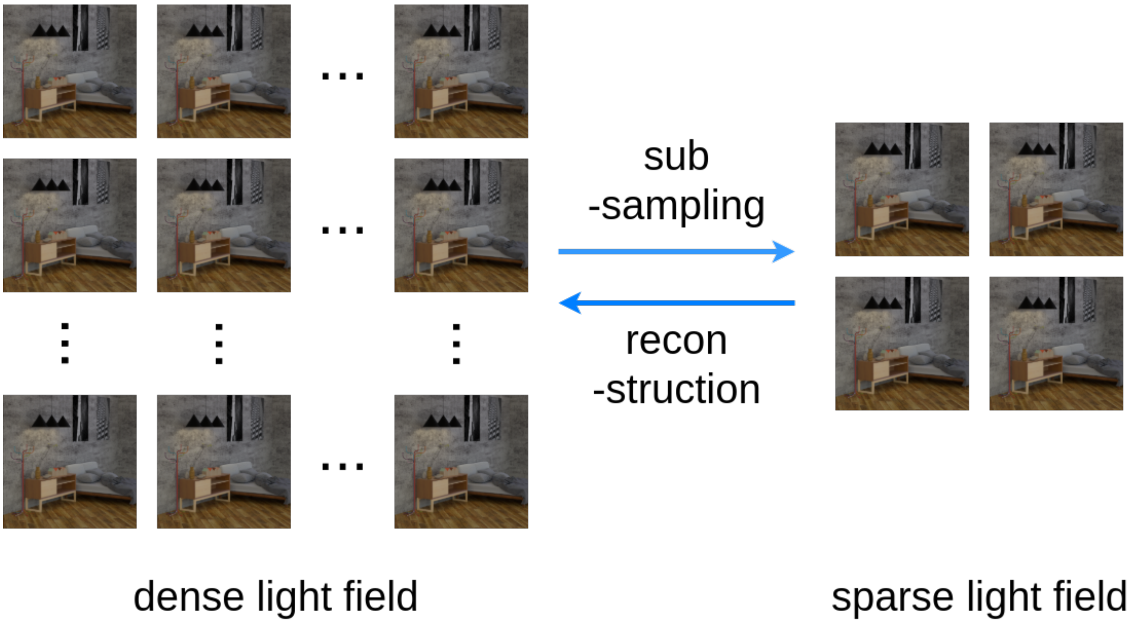


Figure 4.1: Subsampling and reconstruction of a two-plane representation light field.

fixed strategy, such as horizontal & vertical subsampling [108, 111, 115], corners subsampling [80, 101, 103, 105], crosshair subsampling [86, 148] or learned view selection subsampling [90]. To the best of our knowledge, there is no comprehensive scientific investigation comparing different view selection strategies.

Depth based light field view synthesis methods usually start with four input views and focus on generating novel intermediate views on the row, column and central middle positions [80, 101, 103, 105]. Besides, epipolar plane image based methods usually choose the horizontal & vertical strategy and avoid the issues of depth estimation [108, 111, 115]. It is reasonable that epipolar plane image based method rely on this type of strategy as constructing an epipolar plane image requires at least three views along one angular dimension. Besides the view synthesis, view selection is important for other vision tasks on light fields, such as depth estimation and spatial super-resolution. Crosshair strategy is chosen by recent studies as their view selection strategy, as it provides comprehensive epipolar geometry along different angular directions [86, 148]. More recently, learning-based view selection is proposed by Tsai et al. to refine depth estimation on the light field [90]. This method involves all views and determines the importance of each view via an attention-based selection system. The redundancy among sub-aperture views can be reduced by such a system.

4.1.2 Video Frame Interpolation for Light Field View Synthesis

Existing video frame interpolation methods can benefit the light field angular super-resolution problem, as they have a similar configuration. Liu et al. employ an end-to-end fully-convolutional deep network, Deep Voxel Flow, to guide the video frame interpolation with realistic results [149]. Niklaus et al. proposed to estimate motion and color interpolation within one stage using an adaptive 2D kernel which is estimated from a trained convolutional neural network [150]. However, 2D kernel estimation requires huge memory to store information for all pixels and this shortcoming is addressed by replacing the 2D kernel with two separable 1D convolutional kernels [151]. This work was further improved in [152] by using a so-called context map obtained from the layer response of an existing network, combined with a synthesis network based on the GridNet architecture. To generate multiple frames at different time locations instead of only the middle frame, Jiang et al. proposed an indirect video interpolation method which predicts optical flow first and then warps pixels to obtain the target frames [153]. More recently, a depth-aware video frame interpolation method has been proposed [154] which combines the strength of previous methods by estimating the flow, the context, and the kernel altogether, and in addition, the depth. Each component is estimated using a CNN, and the final frame is synthesized with a CNN using all components as input. Video interpolation has also been applied to light field view synthesis in [155], using fully supervised fine-tuning with conventional loss functions on a small light field dataset.

4.2 Study of Efficient Subsampling and Reconstruction Strategies

In this section, we investigate different strategies for light field subsampling and reconstruction. Firstly, a benchmark light field view interpolation method has to be selected from state-of-the-art approaches to evaluate all strategies. Next, given a fixed sampling ratio, three light field subsampling strategies are studied to reconstruct full-size light fields from each sampled light field (Figure 4.2). Finally, six different reconstruction strategies are explored to generate a dense light field from inputs of varying sparsity (Figures 4.3 & 4.4).

4.2.1 Benchmark Method Selection

As our goal is to investigate sampling and reconstruction strategies for light fields, we first select a benchmark method to be applied in our experiments. The benchmark

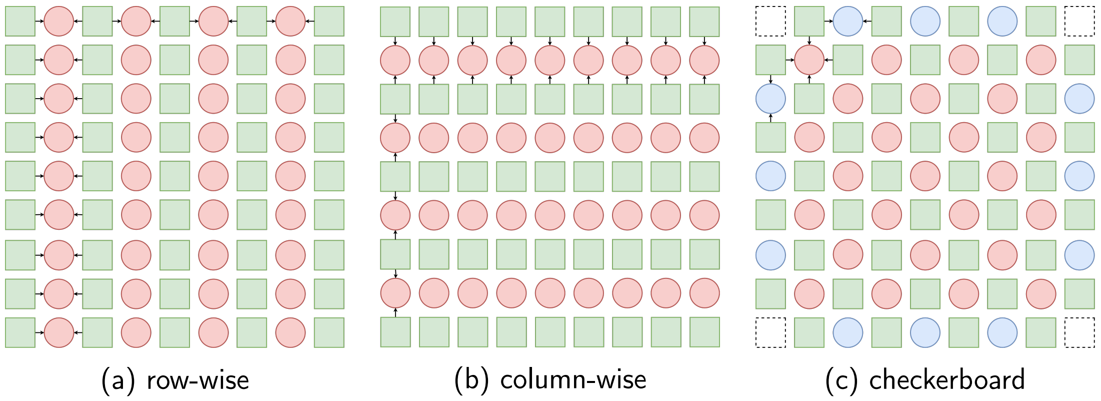


Figure 4.2: Three basic subsampling strategies. The squares are sampled views, the circles are reconstructed views and the dashed squares are unused views. Blue circles in (c) are reconstructed by row-wise or column-wise from two adjacent views to complete the light field.

has to be flexible to work in different configurations but should also provide the best possible interpolation quality. We therefore evaluate SepConv [102], Shearlet [115] and LFEPICNN [156] in an initial study. LFEPICNN is a representative learning-based light field view synthesis method, and Shearlet is an efficient non-learning based reconstruction method in the Fourier domain. SepConv [102] was initially designed for video frame interpolation and employed a neural network-based kernel estimator to interpolate views between adjacent input views. As such it is very flexible and can also be used in various ways of light field view interpolation.

These state-of-the-art methods are evaluated by using the same input pattern shown in Figure 4.2a resulting from row-wise sampling, in which sampled input views are represented as green squares and reconstructed output views are represented as red circles. According to Table 4.1, SepConv numerically outperforms all other methods significantly. Shearlet achieves better performance than linear interpolation. LFEPICNN scores well on PSNR, while Shearlet performs better in terms of average SSIM. However, both these two methods require an epipolar plane image as input. Thus SepConv is not only the best performing approach but also the only one that can easily be adapted to different configurations, as it only needs a pair of RGB images as input. We therefore continue to use SepConv in our further experiments.

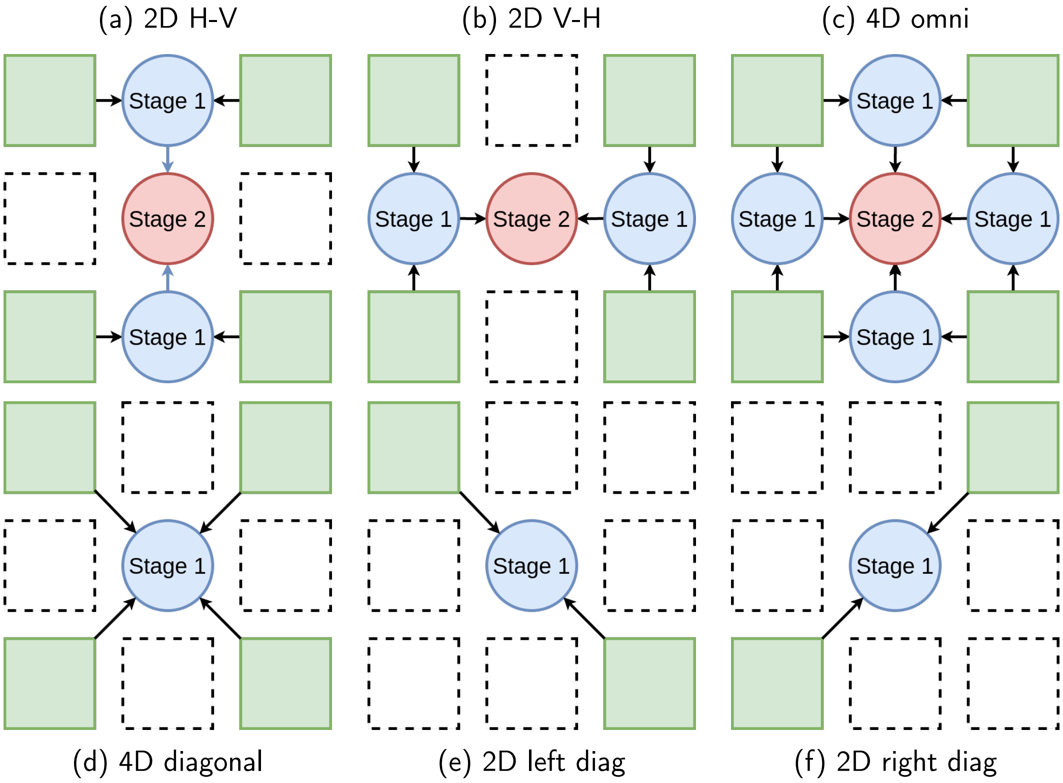
Since SepConv was originally trained for video frame interpolation, we additionally fine-tuned the pre-trained model on our light field training dataset in order to further improve the performance. Table 4.2 shows improvements we can achieve by fine-tuning and retraining the initial network. For some of our experiments detailed below, we have to retrain SepConv appropriately in order to work with more than 2 input views, e.g. left, right, top and bottom neighbors.

Table 4.1: Comparison between state-of-the-art light field interpolation methods

row-wise	PSNR/SSIM		
	Mean	HCI	Stanford
Bilinear	32.37/0.9464	31.82/0.9501	33.47/0.9390
Shearlet [115]	34.92/0.9592	35.78/0.9750	33.21/0.9277
LFEPICNN [156]	37.39/0.9451	37.51/0.9420	37.15/0.9515
SepConv [102]	**38.94/0.9910**	**39.38/0.9945**	**38.07/0.9839**

Table 4.2: Comparison of pretrained, fine-tuned and retrained SepConv for row-wise interpolation

row-wise	pretrained	Fine-tuned	Retrained
PSNR(dB)	38.08	**39.74**	39.41
SSIM	0.9893	**0.9925**	0.9923



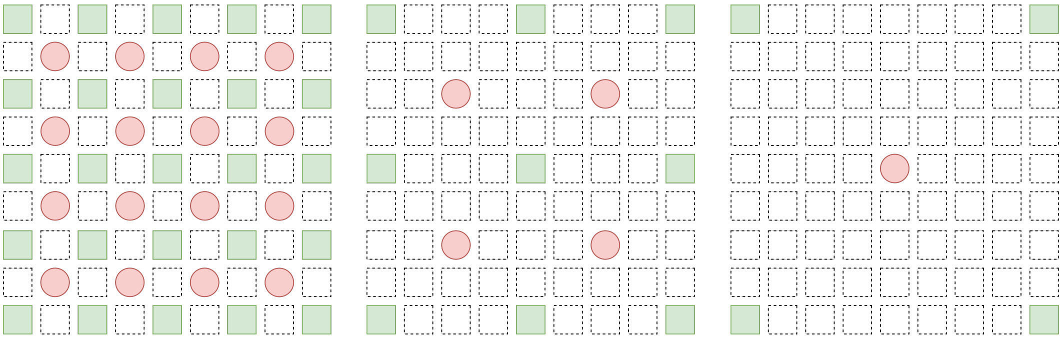
(a) 2D H-V (b) 2D V-H (c) 4D omni

(d) 4D diagonal (e) 2D left diag (f) 2D right diag

Figure 4.3: Six reconstruction strategies to interpolate 3x3 views from 4 input corner views (green), dashed square views are not used for interpolation, different colors of circles identify output views from different stages

4.2.2 Study of Basic Light Field Subsampling Strategies

In this set of experiments, we compare basic subsampling strategies as illustrated in Figure 4.2, with the goal to identify the most efficient basic subsampling strategy among row-wise, column-wise and checkerboard. To evaluate these strategies, the sampled light fields are reconstructed using view interpolation. After reconstruction, the quality of the



(a) level 1 (b) level 2 (c) level 3

Figure 4.4: Three levels of angular density for light field reconstruction. The *IR*s of these levels are 30.9%, 11.1% and 4.9% respectively.

syntesized views can be compared to the corresponding ground truth. All these strategies have the same ratio of sampled views to total views, which is an important measure of the sparsity and defined as the *InputRatio* (*IR*) in equation (4.1):

$$IR = \frac{N_{InputViews}}{N_{InputViews} + N_{OutputViews}} \tag{4.1}$$

where the numbers of input views and output views of the interpolation method are $N_{InputViews}$ and $N_{OutputViews}$, respectively. The total number of views of the completed light field can be represented as the sum of $N_{InputViews}$ and $N_{OutputViews}$. The three basic subsampling strategies, as shown in Figure 4.2 all have $IR \approx 55\%$.

We applied our fine-tuned SepConv, as explained in Section 4.2.1 to interpolate the necessary views for row-wise and column-wise strategies from neighbouring views. For the checkerboard pattern, we had to modify the original SepConv network in order to accept 4 views as input, top, bottom, left and right. The outer views depicted as blue circles in Figure 4.2c were synthesized by row-wise or column-wise interpolation from 2 neighbouring views.

4.2.3 Study of Sparse Light Field Reconstruction Strategies

In this set of experiments, we compare different reconstruction strategies for sparse light fields. Taking a 3x3 matrix of views as an example, six progressive reconstruction strategies can be applied as presented in Figure 4.3. Using four corner images as input, we can reconstruct the side images using the same row-wise and column-wise interpolation as before. Thus, the question becomes, which is the best way to reconstruct the central view. Three of these strategies, including 2D horizontal-vertical (2D H-V), 2D vertical-horizontal (2D V-H) and 4D omni, are two-stage cases involving generating side views as intermediate stage. The other three, including 4D diagonal, 2D left diagonal and 2D

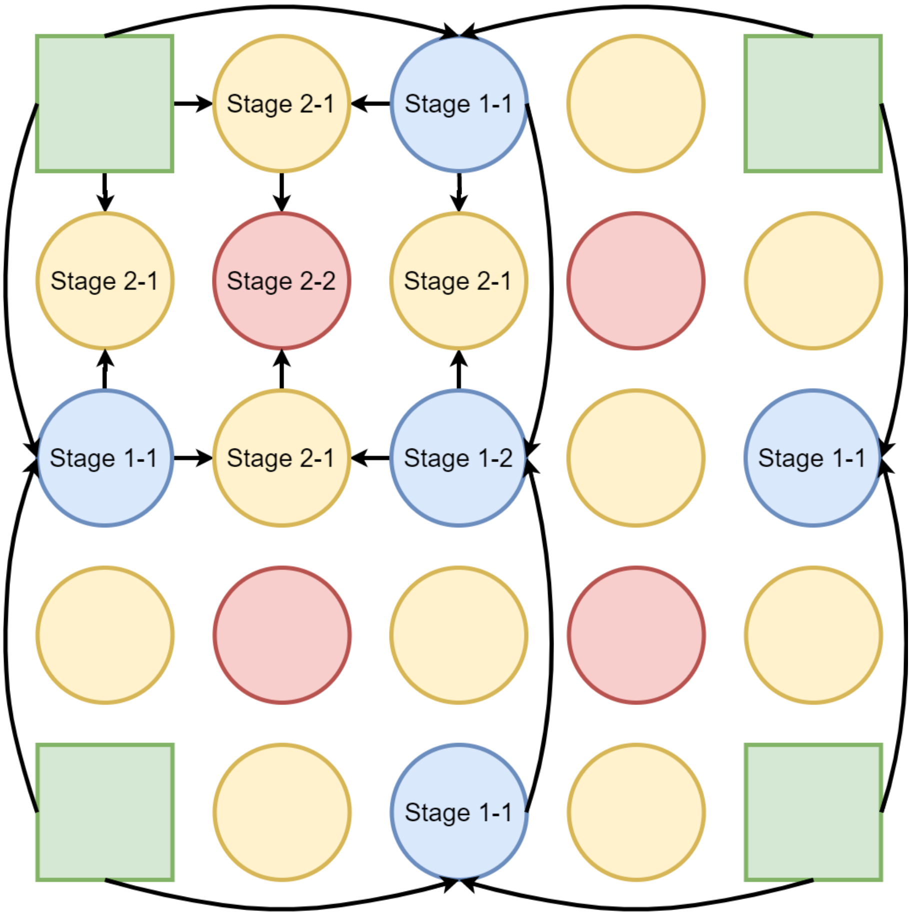


Figure 4.5: Stages to complete a quarter of the full light field for an input angular density of level 2 with $IR = 11.1\%$ (see Fig. 4.4)

right diagonal, require only one stage to generate the central view.

To further study the influence of the angular density of the input views, we investigate three levels of density, as shown in Figure 4.4. The distance between two input views is 1 view in level 1, and there are 3 views and 7 views distance in level 2 and 3, respectively. To compare fully reconstructed light fields, we reconstruct all missing views using appropriate strategies as follows. Level 1 is completed by row-wise and column-wise interpolation of the side views, using the 2D H-V method unless otherwise specified. The choice of 2D H-V as the default method is justified by its better performance demonstrated in Section 4.3. For level 2 and 3, we recursively fill using lower-level methods to complete the full-scale light field. The reconstruction of a quarter of the light field is shown in Figure 4.5, which is applied iteratively to each quarter one by one.

4.3 Results and Evaluation

In this section, we summarize the results of our experiments, which were performed on an Intel Core i7-6700k 4.0GHz CPU, while neural network refining was performed on Nvidia Titan Xp GPUs.

The performance of strategies is evaluated on both real-world and synthetic light field datasets to validate their robustness. We used 27 real-world light fields captured by Lytro Illum cameras provided by EPFL [1] and INRIA [2], and 11 light fields from the Stanford dataset taken by a camera gantry [4]. As for the synthetic light field dataset, all 28 light fields from the HCI benchmark [3] were used. 10 light fields in total were selected from these datasets as the test set and the rest as the training set. Additionally, 160 light fields from light field intrinsic [157] were added for retraining of SepConv to avoid the overfitting. All views were cropped to equal 512x512 resolution to accelerate the computation. In this study, we use the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) on RGB images to evaluate the algorithms numerically.

Table 4.3: Evaluation of three basic subsampling strategies from Figure 4.2

	row-wise		column-wise		checkerboard	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
HCI	40.49	0.9956	40.33	0.9958	39.24	0.9935
Lytro	39.32	0.9932	39.07	0.9925	38.67	0.9921
Stanford	40.22	0.9936	39.37	0.9924	39.47	0.9928
Mean	**39.74**	**0.9925**	39.20	0.9915	39.33	0.9918

Table 4.4: Evaluation of six reconstruction strategies from Figure 4.3

	PSNR	SSIM
2D H-V	**37.42**	**0.9884**
2D V-H	37.21	0.9881
2D left diagonal	35.77	0.9846
2D right diagonal	35.86	0.9842
4D omni	36.51	0.9863
4D diagonal	36.86	0.9838

The results of the evaluation of basic subsampling strategies are shown in Table 4.3. Row-wise interpolation achieves the best scores on most light fields over all datasets. The difference to column-wise is most prominent for the Stanford data which was captured by a gantry, resulting in unequal vertical displacement. For the checkerboard pattern, the retraining of SepConv to accept 4 views may be the reason for its worse performance, as this way it could not benefit from the pre-trained model of SepConv.

Table 4.5: Comparison between level 1 & 2 & 3 from Figure 4.4 using 2D H-V

2D H-V	Level 1	Level 2	Level 3
PSNR(dB)	37.34	35.12	31.79
SSIM	0.9886	0.9825	0.9613

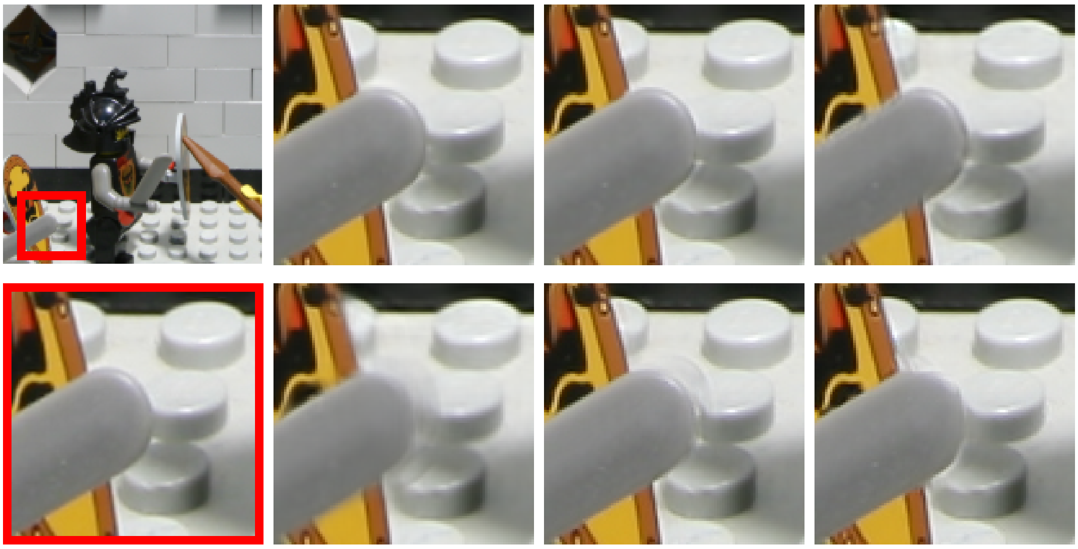
The results of sparse light field reconstruction strategies are shown in Table 4.4. Again, 4D strategies use a retrained model while others use a fine-tuned model. The central view is used to evaluate these results as it is the only one that gets always reconstructed using any of the six strategies when filling a 3x3 block of views. 2D H-V performs best, as it accumulates less error than other strategies with 2 stages (row-wise first gives the best reference for stage 2), while it has a smaller distance between input views compared to direct diagonal strategies. Visual results of reconstruction strategies are shown in Figure 4.6. Occlusion artefacts around the tip of the sword can be observed in diagonal strategies.

Finally, the effect of different levels of angular density is studied in Table 4.5. Since 2D H-V is the optimal strategy according to the previous conclusion, it is utilized to complete the full-scale light fields recursively, as explained before. All synthesized views are averaged in this evaluation (different from only central views in Table 4.4). From Table 4.5 we get the expected decrease of interpolation quality with sparsity.

These insights about sparsity vs quality and the best subsampling and reconstruction strategies can be beneficial for the design of light field coding approaches (the maximum quality that can be achieved when omitting views) or camera systems (maximum camera distance for a desired quality and density).

4.4 Conclusion

In this chapter, we presented a comprehensive study comparing different strategies for efficient light field subsampling and reconstruction. For this purpose, we selected an existing view synthesis method among the best performing state-of-the-art techniques. Using this benchmark method, we first evaluate the best subsampling approach among a row-wise, a column-wise, and a checkerboard pattern with a fixed interpolation ratio, which concludes that the row-wise approach offers the best performances. Second, we investigate corner-based central view generation and compare the performance of six possible reconstruction strategies. In addition, we evaluate a multi-stage approach to reconstruct dense light fields from subsampled input with different levels of angular density. We found that using a row-wise followed by a col-wise reconstruction yields the best performance. We hope these findings will help inspire researches related to light-field subsampling and reconstruction, such as compression and camera array design.



(a) full size GT (b) 2D H-V (c) 2D V-H (d) 4D omni

(e) GT (f) 4D diagonal (g) left diag (h) right diag

Figure 4.6: Visual results of six reconstruction strategies from Figure 4.3

Further, a more explicit analysis regarding the relation of the disparity range to the strategy selection could be performed.

5 Self-supervised Light Field Reconstruction with Cycle Consistency

In this chapter, we propose a self-supervised light field view synthesis framework with cycle consistency. High angular resolution is advantageous for practical applications of light fields. In order to enhance the angular resolution of light fields, view synthesis methods can be utilized to generate dense intermediate views from sparse light field input. Most successful view synthesis methods are learning-based approaches which require a large amount of training data paired with ground truth. However, collecting such large datasets for light fields is challenging compared to natural images or videos. More background details are described in section 2.3. The proposed method aims to transfer prior knowledge learned from high-quality natural video datasets to the light field view synthesis task, which reduces the need for labelled light field data. A cycle consistency constraint is used to build bidirectional mapping enforcing the generated views to be consistent with the input views. Derived from this key concept, two-loss functions, cycle loss and reconstruction loss, are used to fine-tune the pre-trained model of a state-of-the-art video interpolation method. The proposed method is evaluated on various datasets to validate its robustness, and results show it not only achieves competitive performance compared to supervised fine-tuning but also outperforms state-of-the-art light field view synthesis methods, especially when generating multiple intermediate views. Besides, our generic light field view synthesis framework can be adapted to any pre-trained model for advanced video interpolation.

5.1 Related Work

5.1.1 Cycle Consistency

The key element of our proposed method is the introduction of the cycle consistency to the light field angular dimension. The cycle consistency constraint aims to regularize structured predictions between two or more different domains. More recently, this constraint has become a commonly used technique in computer vision area and has been explored on numerous tasks, including image co-segmentation [158, 159], image matching [160, 161, 162], image captioning [163], depth estimation [164], structure from motion [165, 166] and image translation [167]. Beyond the vision area, cycle consistency is also applied in an inverse translation method for the language translation task [168].

The cyclic generation procedure of our work is inspired by the success of cyclic image generation for video interpolation [169, 170]. Cycle consistency constraint is employed as a loss function computing the Euclidean distance between cyclic generated results and original inputs. Different from unidirectional mapping built by conventional cost measurement functions, cycle consistency loss establishes a bidirectional mapping between two spaces. Besides, past researches also demonstrate the strength of cycle consistency to adapt a pre-trained model to a new target domain, i.e. video interpolation to light field angular reconstruction. For instance, Gao and Koch extend pre-trained video interpolation model to the light field reconstruction by proposing a parallax-interpolation adaptive separable convolution [155]. Similar to our work, Gao et al. [171] also utilize cycle consistency to the light field reconstruction task. They process light field as the sparsely regularized epipolar plane image in the shearlet domain, while our work stays in sub-aperture space that can be adapted to unaligned views, such as unstructured light field captured by single-lens cameras.

5.1.2 Self-supervised Learning

Although strongly supervised learning demonstrates successfully applications on various visual benchmarks [172, 173], such success requires numerous annotated data, which is not always available as data collection and manual labelling could be expensive. As a subclass of the typical unsupervised learning methods, self-supervised learning aims to solve this problem by learning the representation from the unlabelled data. In other words, the learning process can be understood as it is guided by the surrogate "pretext" tasks, i.e. automated generated labels, high-level representation or input data directly, that can be modelled from only unsupervised data. So far, self-supervised methods produce a superior performance on a broad range of challenging tasks [174, 175, 176].

For instance, Doersch et al. introduce a typical self-supervised framework by exploring the use of relative positioning of cropped patches as the spatial context [177]. The key hypothesis of this work is that the high-level representation of context is properly modelled while solving this spatial relationship problem as a "pretext" task. Therefore, those learned representation embedded in the networks can benefit other downstream tasks such as object detection and data mining. Follow-up researches generalize this framework by estimating the permutation of multiple patches sampled and permuted from images [178, 179]. Besides, Zhai et al. aim to properly model a special type of dataset containing a small amount of labelled data along with a huge amount of unlabelled data [180]. To address this problem, they propose a new learning technique by bridging semi-supervised learning with self-supervised learning, as a small amount of labelled data would benefit self-supervised representation learning. Adapting similar self-supervised structure to the video interpolation work, Reda et al. introduce the pseudo supervision to avoid the use of the intermediate frame as the ground truth [170]. For a comprehensive review of previous self-supervised learning, we refer the reader to read these papers [181, 182].

Light field data acquisition is still a tedious process and even largest light field dataset, which only has $\sim$ 3300 light field images [100], is relatively small compared to commonly available single image and video datasets, such as ImageNet($\sim$ 14 million images) [183], COCO($>$ 1 million images) [184] and Aff-Wild($\sim$ 1 million images) [185]. Without sufficient support of ground truth supervision, light field reconstruction becomes a suitable problem to be addressed with the self-supervised learning technique. Furthermore, self-supervised learning can be motivated by the cycle consistency on the light field angular domain, which is trained solely on the sparse light fields. Moreover, we also consider transferring the prior knowledge from video interpolation task to light field reconstruction task, while only requiring a limited amount of light field data as fine-tuning datasets. The closest work utilizing cycle consistency to enable the self-supervised training is proposed in the shearlet domain [171], while our work concentrates on the sub-aperture domain.

5.2 Light Field View Synthesis using Cycle Consistency

5.2.1 Problem Formulation

In this work, a 4D light field $\mathbf{L}$ is parameterized using the two parallel planes representation as depicted in Figure 5.1, indexed by x, y over the spatial dimensions and s, t are the angular dimensions. We denote by $\mathbf{I}_{s,t}$ the view extracted from a light field $\mathbf{L}$ at the angular position s, t. Given a sparsely-sampled light field $\mathbf{L}^S$ with resolution $(H \times W \times n \times n)$, the goal is to reconstruct a more densely-sampled light field $\mathbf{L}^D$ with the same spatial resolution and a higher angular resolution $(H \times W \times N \times N)$, where $N = \alpha(n-1)+1$ and α is the up-sampling factor in the angular domain. Unless mentioned specifically, $\alpha = 2$ is used as default to explain our method. By fixing one angular

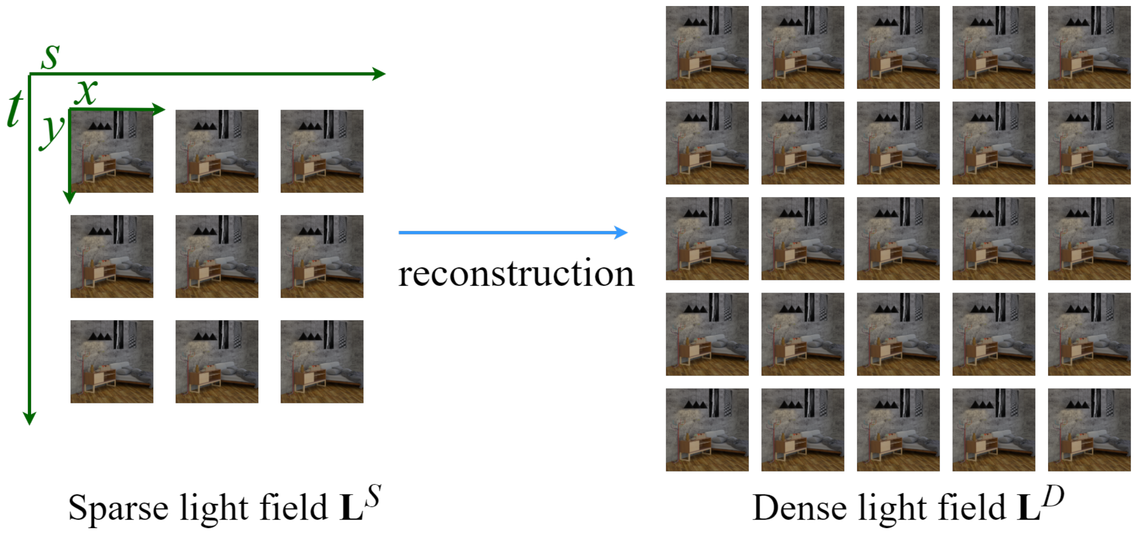

s
t
x
y
reconstruction
Sparse light field $\mathbf{L}^S$
Dense light field $\mathbf{L}^D$

Figure 5.1: Dense light field reconstruction. We aim to reconstruct a dense light field L^D with angular resolution (N, N) from a sparse light field L^S with angular resolution (n, n). The spatial resolution (H, W) of each view remains unchanged.

dimension, a set of views can be extracted along the remaining angular dimension of the light field. Such a view set can be considered as a consecutive frame sequence, which can be captured by a virtual camera moving along the corresponded angular direction. Thus, the dense light field reconstruction problem can be treated as a video interpolation process along the fixed angular dimension. Many CNN-based methods have been shown to be successful for video interpolation tasks. However, directly adopting a pre-trained network from an existing video interpolation method to the light field domain may fail since the distribution of these two kinds of data may differ. On the other hand, retraining a CNN from scratch can be laborious and the limited size of light field datasets may not allow reaching competitive performance. Thus, to maximally leverage the advantage of the cutting-edge video interpolation methods and to avoid its troublesome retraining, we introduce a self-supervised fine-tuning approach using cycle consistency to apply the pre-trained model of a video interpolation method to the light field domain.

5.2.2 Proposed Framework with Self-Supervised Learning

Given a sparse light field $\mathbf{L}^S$ with angular resolution (n, n), our proposed approach aims to build a learning-based model that takes this light field as input and accurately reconstructs a high-quality dense light field $\mathbf{L}^D$ with angular resolution (N, N) without the support of paired ground truth. As shown in Figure 5.2, we consider triplets of views extracted from $\mathbf{L}^S$ along a fixed angular dimension, either horizontally $\{\mathbf{I}^S_{s-2,t}, \mathbf{I}^S_{s,t}, \mathbf{I}^S_{s+2,t}\}$ or vertically $\{\mathbf{I}^S_{s,t-2}, \mathbf{I}^S_{s,t}, \mathbf{I}^S_{s,t+2}\}$. Note that triplets have to be used due to our proposed cycle loss described below. A dense light field $\mathbf{L}^D$ is obtained by first performing horizontal interpolation on all rows, and then performing vertical interpolation on all columns.

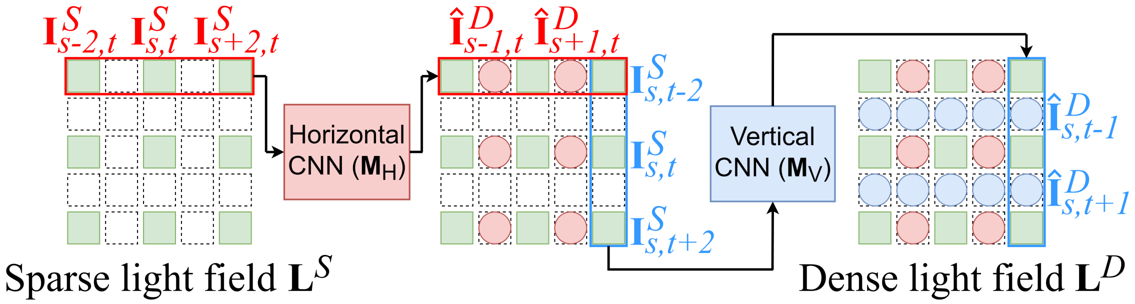

$\mathbf{I}^S_{s\text{-}2,t}$ $\mathbf{I}^S_{s,t}$ $\mathbf{I}^S_{s+2,t}$
Horizontal CNN ($\mathbf{M}_\mathrm{H}$)
$\hat{\mathbf{I}}^D_{s\text{-}1,t}$ $\hat{\mathbf{I}}^D_{s+1,t}$
$\mathbf{I}^S_{s,t\text{-}2}$
$\mathbf{I}^S_{s,t}$
$\mathbf{I}^S_{s,t+2}$
Vertical CNN ($\mathbf{M}_\mathrm{V}$)
$\hat{\mathbf{I}}^D_{s,t\text{-}1}$
$\hat{\mathbf{I}}^D_{s,t+1}$
Sparse light field $\mathbf{L}^S$
Dense light field $\mathbf{L}^D$

Figure 5.2: An overview of our proposed view interpolation approach. Horizontal and vertical interpolation are cascaded to reconstruct L^D from L^S using view triplets from the corresponding angular dimensions.

The view interpolation is achieved by two CNNs which share the same architecture but are trained separately along the horizontal and vertical dimensions.

Let us consider the horizontal interpolation case in order to explain the framework more in detail. Given an input triplet $\{\mathbf{I}^S_{s-2,t}, \mathbf{I}^S_{s,t}, \mathbf{I}^S_{s+2,t}\}$, two intermediate views can be generated from pairwise adjacent views:

$$\begin{aligned}\hat{\mathbf{I}}^D_{s-1,t} &= \mathbf{M}(\mathbf{I}^S_{s-2,t}, \mathbf{I}^S_{s,t})\\ \hat{\mathbf{I}}^D_{s+1,t} &= \mathbf{M}(\mathbf{I}^S_{s,t}, \mathbf{I}^S_{s+2,t})\end{aligned} \tag{5.1}$$

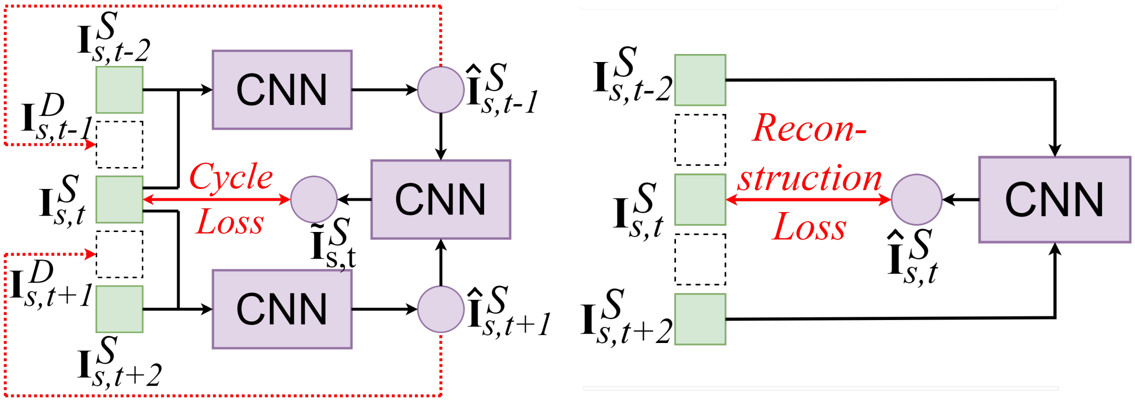
where $\mathbf{M}$ is a pre-trained video interpolation method.

Inspired by the recent success of the application of the cycle consistency for video interpolation [169, 170], we propose to fine-tune our baseline interpolator $\mathbf{M}$ in a self-supervised manner by applying the cycle consistency constraint to the light field angular domain, as shown in Figure 5.3a. By applying the interpolator $\mathbf{M}$ on the two intermediate views generated from the input triplet as defined in equation 5.1, we can obtain an estimate of the center view of the input triplet $\mathbf{I}^S_{s,t}$ which we denote as the cycle-reconstructed view $\tilde{\mathbf{I}}^S_{s,t}$:

$$\begin{aligned}\tilde{\mathbf{I}}^S_{s,t} &= \mathbf{M}\left(\hat{\mathbf{I}}^D_{s-1,t},\ \hat{\mathbf{I}}^D_{s+1,t}\right)\\ &= \mathbf{M}\left(\mathbf{M}(\mathbf{I}^S_{s-2,t}, \mathbf{I}^S_{s,t}),\ \mathbf{M}(\mathbf{I}^S_{s,t}, \mathbf{I}^S_{s+2,t})\right)\end{aligned} \tag{5.2}$$

We can thus define the cycle-loss as the ℓ_1-norm distance between the cycle-reconstructed view $\hat{\mathbf{I}}^S_{s,t}$ and the input view $\mathbf{I}^S_{s,t}$:

$$\mathcal{L}_c = ||\tilde{\mathbf{I}}^S_{s,t} - \mathbf{I}^S_{s,t}||_1 \tag{5.3}$$


$\mathbf{I}^S_{s,t\text{-}2}$
$\mathbf{I}^D_{s,t\text{-}1}$
CNN
$\hat{\mathbf{I}}^S_{s,t\text{-}1}$
$\mathbf{I}^S_{s,t}$
Cycle
Loss
CNN
$\tilde{\mathbf{I}}^S_{s,t}$
$\mathbf{I}^D_{s,t+1}$
CNN
$\hat{\mathbf{I}}^S_{s,t+1}$
$\mathbf{I}^S_{s,t+2}$
$\mathbf{I}^S_{s,t\text{-}2}$
Recon-
struction
Loss
CNN
$\mathbf{I}^S_{s,t}$
$\hat{\mathbf{I}}^S_{s,t}$
$\mathbf{I}^S_{s,t+2}$

(a) Cycle loss

(b) Reconstruction loss

Figure 5.3: Illustration of the cycle loss and reconstruction loss on a vertical input triplet. Both losses do not require any knowledge of the ground truth intermediate views (represented by dashed square) and can therefore be used for self-supervised training.

While ℓ_1-norm based losses are able to minimize the overall error between the estimated images and the corresponding original images, they are known to generate over-smooth results. To tackle this problem, we also introduce in our framework a perceptual loss $\mathcal{L}_p$, defined as the ℓ_2-norm between high-level convolutional features extracted from the cycle-reconstructed view and the input view:

$$\mathcal{L}_p = ||\Psi(\tilde{\mathbf{I}}^S_{s,t}) - \Psi(\mathbf{I}^S_{s,t})||_2 \tag{5.4}$$

where Ψ extracts the convolutional features from images using a VGG-16 network [106], which then is applied to train our base CNN network (SepConv [151], see below).

Furthermore, to stabilize the training process, we introduce a reconstruction loss of $\mathcal{L}_r$, as shown in Figure 5.3b. In this case, the two non-adjacent views from the input triplet $\mathbf{I}^S_{s-2,t}$ and $\mathbf{I}^S_{s+2,t}$ are used to generate the center view of the input triplet:

$$\hat{\mathbf{I}}^S_{s,t} = \mathbf{M}\left(\mathbf{I}^S_{s-2,t}, \mathbf{I}^S_{s+2,t}\right) \tag{5.5}$$

This reconstructed view can be used to define the reconstruction loss $\mathcal{L}_r$ as its ℓ_1-norm distance to the input view:

$$\mathcal{L}_r = ||\hat{\mathbf{I}}^S_{s,t} - \mathbf{I}^S_{s,t}||_1 \tag{5.6}$$

Note that all losses introduced in our framework as defined in equations 5.3, 5.4, and 5.6, do not rely on any knowledge of the ground truth dense light field $\mathbf{L}^D$ but only the given sparse input light field $\mathbf{L}^S$, thus allowing to perform self-supervised training or fine-tuning of the learning-based interpolator $\mathbf{M}$.

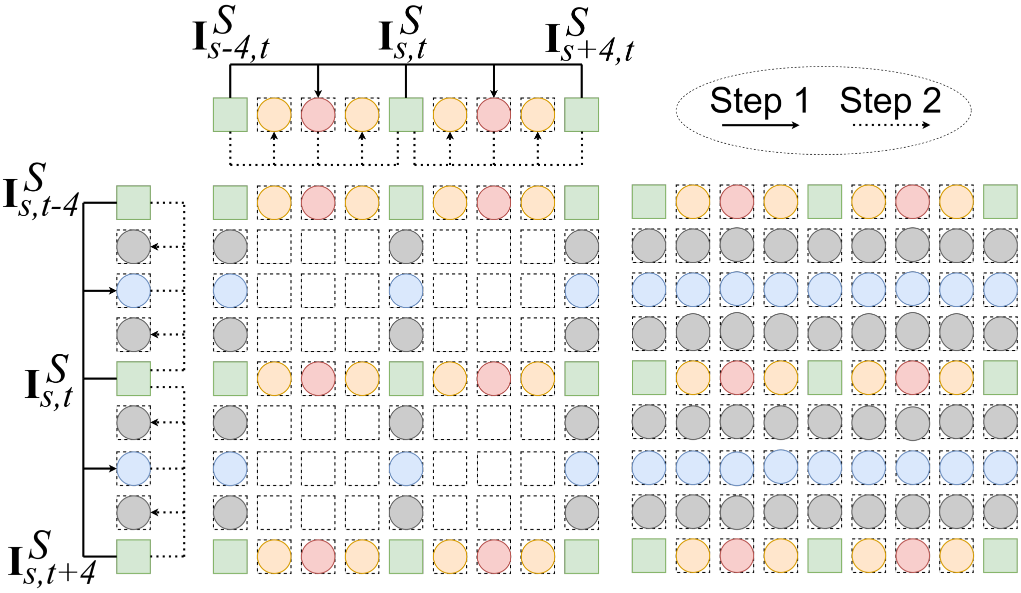


Figure 5.4: Demonstration of the two-step strategy to generate multiple intermediate views from a set of sparse views, denoted as green squares, when $\alpha = 4$. The first step is to synthesize middle views, denoted as red and blue circles, between each pairwise input views. The second step uses original and synthetic views to reconstruct the remaining missing views, denoted as yellow and grey circles, along one angular dimension.

Multi-step Light Field Generation While our framework naturally performs angular up-sampling with a factor $\alpha = 2$, denser light fields can be obtained by iteratively applying the proposed approach, as illustrated in Figure 5.4 for $\alpha = 4$. Any upsampling factor which is a power of two is in fact supported, *i.e.* $\alpha = 2^x, (x \in \mathbb{Z}\ \&\ x > 1)$.

5.2.3 Implementation Details

In this work, we select the adaptive separable convolution (SepConv) [151] as our baseline interpolator **M** due to its balance between ease of use and performance accuracy, but note that any learning-based video interpolation method [152, 153, 154] can be used within our framework. The network of SepConv employs an encoder-decoder architecture, each part contains convolution blocks and skip connections, to extract features and then performs four 1D kernel estimations individually to obtain the final results. We use the implementation available online based on *PyTorch* [12] and use the default configurations from the original SepConv paper. We fine-tune the pre-trained model by minimizing the objective function:

$$\arg\min_{\mathbf{M}} \left(\lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_p \mathcal{L}_p\right) \tag{5.7}$$

[1] github.com/sniklaus/sepconv-slomo

[2] github.com/HyeongminLEE/pytorch-sepconv

where $\mathcal{L}_c$, $\mathcal{L}_r$ and $\mathcal{L}_p$ are defined in equations (5.3), (5.6) and (5.4). For all experiments, we set the parameters as $\lambda_c = 1$, $\lambda_r = 1$, and $\lambda_p = 0.06$. The Adam optimizer is applied for optimization with a batch size of 8. We start with the learning rate of 0.001 and a scheduler is applied to decay the rate according to the learning progress. As in the original SepConv work, we firstly crop training data to 150×150 patches, then randomly crop to 128×128. In addition, we perform pre-processing to eliminate patches containing too small disparity. An Intel Core i7-6700k 4.0GHz CPU was used for all our experiments, and the neural network training was run on a single Nvidia Titan Xp GPU with 12 GB memory.

5.3 Results and Evaluation

In this section, we first conduct an ablation study, especially evaluating the efficiency of the proposed framework compared to supervised fine-tuning. For this purpose, we use a variety of real-world and synthetic dense light field datasets which we sub-sample to create our test sparse datasets with sampling ratios $\alpha = 2$ and $\alpha = 4$.

We then compare the proposed framework to two representative light field view synthesis methods, a shearlet-based method [115] and a learning-based method (LFEPICNN) [109], along with bilinear interpolation and disparity based view synthesis proposed in Chapter 3.

For all our evaluations, the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) are computed over RGB images to evaluate the numerical performance of the different methods. For each light field, unless emphasized specifically, the average numerical results are computed over all synthesized views. All evaluations are performed on the same machine to ensure the fairness of the comparison.

5.3.1 Ablation Study

For this study, we used dense light fields from real-world and synthetic datasets. For the real-world dataset, we selected 27 real-world Lytro light fields captured by EPFL [1] and INRIA [2] using Lytro Illum cameras, and 11 light fields from the Stanford dataset taken by a camera gantry [4]. The Lytro Illum light fields are processed with the pipeline of Matysiak et al. [186]. For the synthetic light field dataset, all 28 light fields from the HCI benchmark [3] were used, as well as 160 light fields from the dataset of [157].

For testing, 10 light fields are used: 2 from EPFL, 2 from INRIA, 2 from Stanford, and 4 from HCI. All remaining light fields are used for training.

Test sparse light fields are sub-sampled from the original light fields with ratios $\alpha = 2$ and

$\alpha = 4$. More precisely, 9×9 views are extracted from input light fields and considered as dense ground truth, and 5×5 and 3×3 views are then sub-sampled to create sparse light fields.

We conduct the ablation experiments by comparing to several variants of the proposed framework. First, we use the pre-trained model of SepConv as the baseline. Since the dense light field ground truth is available, we fine-tuned the SepConv model using supervised training. We also evaluate the influence of the cycle loss by training our framework using only the reconstruction loss. We also assess the performance of our framework when vertical interpolation is performed before horizontal interpolation, as opposed applying horizontal interpolation first as shown in Figure 5.2.

The numerical results are computed and averaged over all test light fields, and the comparison is presented in Table 5.1. As we can observe, our proposed method can outperform the pre-trained model even without the support of the ground-truth, and achieve competing performance compared to fully supervised fine-tuning. It is also clear that the use of the cycle loss improves the performance of our framework. In addition, we can see that the cascading order of horizontal/vertical or vertical/horizontal interpolation has a non-negligible impact on the final performance.

Table 5.1: Quantitative results of the ablation study.

	$\alpha = 2$		$\alpha = 4$	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
SepConv Pretrained	37.23	0.9880	34.66	0.9793
SepConv Supervised Fine-tuning	38.40	0.9921	35.81	0.9831
Ours without Cycle Loss	38.01	0.9883	35.25	0.9801
Ours with V-H CNN	38.14	0.9889	35.67	0.9817
Ours Full Model	38.30	0.9902	35.72	0.9830

5.3.2 Comparison to Light Field View Synthesis Methods

We compare our proposed framework against two existing light field view synthesis methods: shearlet-based reconstruction(Shearlet) [115], and LFEPICNN [109]. Moreover, we also compared it to the pre-trained SepConv model and the disparity-based view synthesis (DVS) results from our previous approach described in Chapter 3. We used the implementations of Shearlet and LFEPICNN methods provided by original authors, and carefully selected their parameters to maximize their performance.

Synthetic and real-world Datasets. We first perform evaluation on the synthetic HCI, real-world Lytro and Stanford datasets, as each dataset corresponds to a different disparity range discussed in 3.3.2. The quantitative of which is shown in Table 5.3, Table 5.2, and Table 5.4. Please note the each PSNR/SSIM score is average among all synthetic sub-aperture views for each light field.

The shearlet-based reconstruction is almost always outperformed by all other four methods including ours. While shearlet-based reconstruction and LFEPICNN are designed specifically for light fields and DVS mainly relies on the performance of the disparity estimator, they are only competitive on the Lytro dataset which has a very narrow disparity range. Thus except for the Lytro dataset with $\alpha = 2$, our proposed framework consistently outperforms other light field view synthesis methods. Thanks to the application of cycle consistency, our method also outperforms the pre-trained SepConv method. In addition, it can be observed that our method is more robust when using sparser input datasets such as when using a sub-sampling ratio $\alpha = 4$ over all datasets.

We present visual comparisons for the *ChezEdgar* light field from the Lytro dataset and *LegoKnights* from the Stanford dataset as examples of real-world scenes in Figure 5.5. *LegoKnights* is a challenging case as it has wider disparity than other test light fields and large texture-less regions. Shearlet [115], LFEPICNN [109] and DVS methods all fail to produce plausible results and significant artefacts can be observed on challenging areas, such as the tip of the sword and bricks on the background wall. In comparison, our proposed approach generates results closer to the ground-truth. It demonstrates that our method is more robust to different real-world scenes and is able to produce more photo-realistic results for large disparity view synthesis.

A visual comparison of synthetic scenes is presented in Figure 5.6 using the *Herbs* and *Bicycle* light fields from the HCI dataset [3]. As we can observe, Shearlet [115] fails to reconstruct sharp details in texture-less regions, such as the door in *Bicycle*. The results of LFEPICNN [109] are blurry in occluded regions, such as the leaves in *Herbs* and the metal bin in *Bicycle*. To conclude, our method produces promising quantitative and qualitative results on the synthetic HCI dataset and shows robustness to occlusions and texture-less surfaces.

Table 5.2: Numerical results on the real-world Lytro datasets [1, 2]

$\alpha = 2$	PSNR(dB)					SSIM				
	average	Bee_2	Bikes	ChezEdgar	Desktop	average	Bee_2	Bikes	ChezEdgar	Desktop
Shearlet	33.11	32.34	34.32	32.82	32.95	0.9667	0.9467	0.9702	0.9755	0.9744
LFEPICNN	35.35	35.23	36.34	34.40	35.45	0.9864	**0.9760**	0.9928	0.9877	0.9892
SepConv	35.30	32.51	35.71	36.80	36.17	0.9836	0.9602	0.9942	0.9900	0.9901
Bilinear	**37.33**	**35.49**	**37.98**	**38.03**	**37.82**	0.9870	0.9705	0.9932	**0.9925**	**0.9918**
DVS	36.40	35.31	37.60	37.44	35.25	0.9866	0.9702	0.9930	0.9920	0.9912
CycleLF	36.76	34.96	37.26	37.53	37.28	**0.9876**	0.9727	**0.9949**	0.9917	0.9911

$\alpha = 4$	PSNR(dB)					SSIM				
	average	Bee_2	Bikes	ChezEdgar	Desktop	average	Bee_2	Bikes	ChezEdgar	Desktop
Shearlet	29.99	27.89	30.42	30.58	31.09	0.9361	0.8423	0.9592	0.9738	0.9690
LFEPICNN	32.06	31.45	32.85	31.69	32.25	0.9640	0.9264	0.9804	0.9761	0.9732
SepConv	32.46	30.58	32.90	33.62	32.72	0.9712	0.9370	0.9859	0.9822	0.9797
Bilinear	32.60	30.29	32.50	34.12	33.10	0.9753	0.9010	0.9721	0.9835	0.9758
DVS	32.25	30.18	32.53	34.02	32.27	0.9594	0.8998	0.9745	0.9842	0.9793
CycleLF	**33.62**	**32.61**	**33.93**	**34.20**	**33.73**	**0.9767**	**0.9535**	**0.9874**	**0.9845**	**0.9813**

Table 5.3: Numerical results on the synthetic HCI dataset [3]

$\alpha = 2$	PSNR(dB)					SSIM				
	average	bedroom	bicycle	herbs	origami	average	bedroom	bicycle	herbs	origami
Shearlet	34.82	38.10	34.23	30.09	36.86	0.9734	0.9677	0.9864	0.9440	0.9955
LFEPICNN	34.25	36.98	32.65	31.92	35.47	0.9692	0.9856	0.9834	0.9264	0.9814
SepConv	38.89	41.22	36.16	37.06	41.11	0.9943	0.9946	0.9941	0.9911	0.9976
Bilinear	31.45	33.37	30.71	28.76	32.97	0.9397	0.9672	0.9649	0.8586	0.9682
DVS	37.67	41.08	34.80	35.82	38.98	0.9931	0.9945	0.9929	0.9880	0.9972
CycleLF	**39.87**	**41.77**	**37.15**	**38.51**	**42.06**	**0.9953**	**0.9949**	**0.9948**	**0.9937**	**0.9980**

$\alpha = 4$	PSNR(dB)					SSIM				
	average	bedroom	bicycle	herbs	origami	average	bedroom	bicycle	herbs	origami
Shearlet	29.88	31.20	28.88	27.34	32.13	0.8911	0.9246	0.9038	0.8083	0.9274
LFEPICNN	30.42	32.83	29.42	28.43	31.01	0.9172	0.9449	0.9472	0.8257	0.9508
SepConv	36.24	39.43	33.39	34.12	38.01	0.9888	0.9923	0.9875	0.9785	0.9958
Bilinear	28.04	30.09	26.82	26.69	28.55	0.8733	0.9069	0.8909	0.7760	0.9194
DVS	34.94	39.44	31.81	33.15	35.37	0.9855	0.9931	0.9824	0.9732	0.9933
CycleLF	**37.44**	**39.77**	**34.73**	**35.68**	**39.58**	**0.9913**	**0.9932**	**0.9884**	**0.9866**	**0.9970**

Table 5.4: Numerical results on the real-world Stanford Gantry datasets [4]

$\alpha = 2$	PSNR(dB)			SSIM		
	average	LegoKnights	TheStanfordBunny	average	LegoKnights	TheStanfordBunny
Shearlet	31.46	24.79	38.13	0.8906	0.7990	0.9822
LFEPICNN	34.68	29.38	39.97	0.9407	0.8967	0.9846
SepConv	37.80	34.22	41.38	0.9843	0.9797	0.9889
Bilinear	32.57	26.17	38.98	0.9141	0.8463	0.9819
DVS	35.67	30.64	40.70	0.9765	0.9642	0.9888
CycleLF	**38.23**	**34.72**	**41.73**	**0.9853**	**0.9816**	**0.9890**

$\alpha = 4$	PSNR(dB)			SSIM		
	average	LegoKnights	TheStanfordBunny	average	LegoKnights	TheStanfordBunny
Shearlet	29.04	22.59	35.49	0.8484	0.7201	0.9768
LFEPICNN	30.46	24.81	36.11	0.8762	0.8010	0.9514
SepConv	35.92	31.31	40.53	0.9767	0.9651	0.9883
Bilinear	28.60	22.85	34.35	0.8441	0.7398	0.9484
DVS	32.87	27.51	38.23	0.9586	0.9303	0.9869
CycleLF	**36.47**	**31.92**	**41.02**	**0.9791**	**0.9696**	**0.9885**

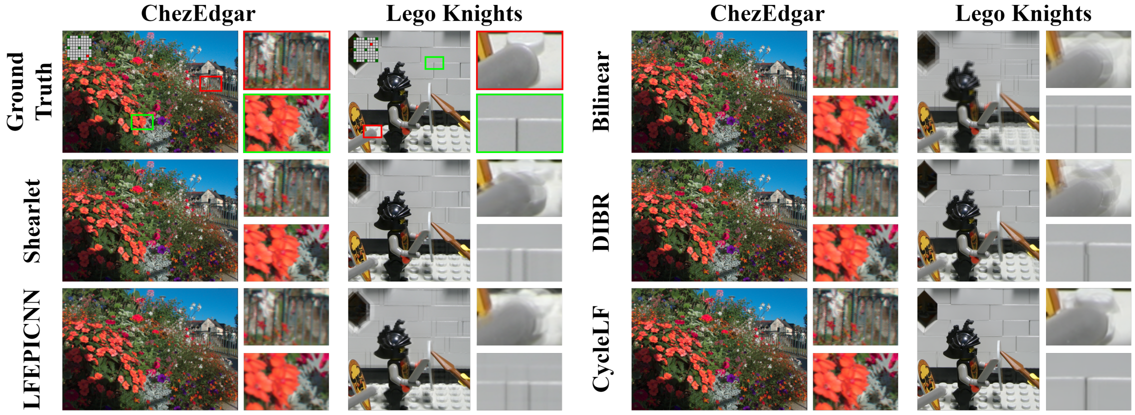


Figure 5.5: Visual comparison on the INRIA ChezEdgar and Stanford Lego Knights light fields. (a) Ground-truth. (b) Shearlet [115]. (c) LFEPICNN [109]. (d) Bilinear. (d) DVS. (e) CycleLF.

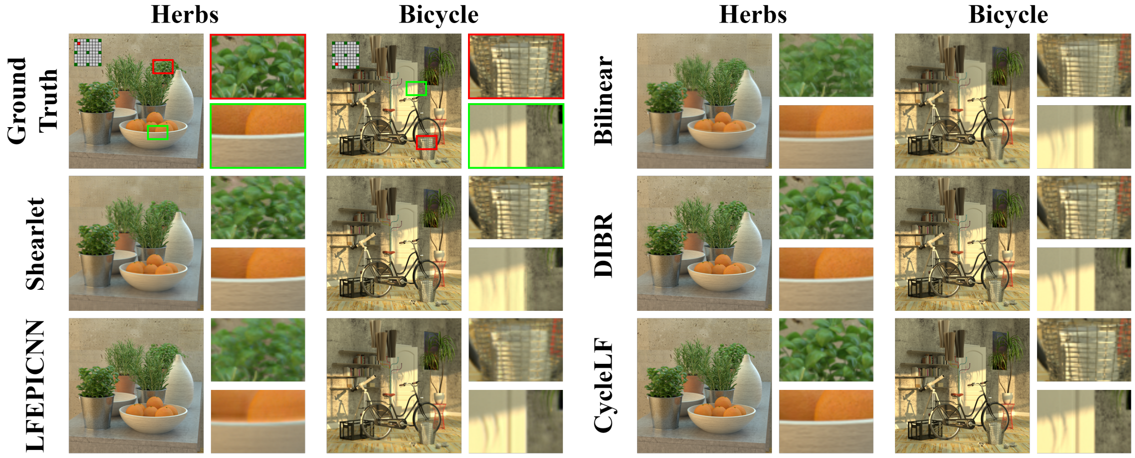


Figure 5.6: Visual comparison on the synthetic HCI dataset [3]. (a) Ground-truth. (b) Shearlet [115]. (c) LFEPICNN [109]. (d) Bilinear. (d) DVS. (e) CycleLF.

Wide Baseline Datasets. To investigate the performance of our view synthesis method on light field scenes with large disparity range, we evaluate our method on three wide baseline datasets: HDCA [5], MPI [6] and CIVIT [7]. We also compare our method to a more recent light field view synthesis method CycleST [171]. The High Density Camera Array (HDCA) dataset [5] is originally composed of size $101(99) \times 21 \times 3976 \times 2652 \times 3$. We followed the same data preprocessing as [171, 187] to unify the resolution as $97 \times 21 \times 1280 \times 720$ and avoid the black border observed in the original images. When the sampling factor α is set as 16, the resulting input number of views is 7. Note that we only perform horizontal view synthesis for every row of light field input views. We also select two light fields from the MPI light field archive [6] and five light fields from the Centre for Immersive Visual Technologies (CIVIT) dataset [7]. Both of them are horizontal-parallax light fields. We keep the same experiment setting as the HDCA dataset, in that the top 97 views are used as ground truth to be reconstructed from input 7 views along with sampling ratio $\alpha = 16$. The numerical results for HDCA,

Table 5.5: Numerical results on the HDCA dataset [5]

$\alpha = 16$	Disparity(pix)			PSNR(dB)		SSIM	
	d_{min}	d_{max}	d_{range}	CycleST	CycleLF	CycleST	CycleLF
average				**31.15**	19.85	**0.9391**	0.6939
Set Books and charts Scene	25.0	44.0	19.0	**34.43**	18.97	**0.9818**	0.8378
Set Lego City Scene	27.0	49.0	22.0	**27.27**	16.18	**0.8871**	0.4361
Set Lightfield Production Scene	28.0	55.0	27.0	**30.44**	18.22	**0.9332**	0.5126
Set Plants Scene	25.0	55.0	30.0	**32.01**	19.04	**0.9522**	0.6999
Set Table in the garden Scene	25.0	54.0	29.0	**32.74**	23.21	**0.9551**	0.8526
Set TableTop I Scene	25.0	54.0	29.0	**36.24**	23.51	**0.9767**	0.8245
Set TableTop II Scene	26.0	43.0	17.0	**27.47**	16.50	**0.9107**	0.5373
Set TableTop III Scene	28.0	49.0	21.0	**28.00**	16.66	**0.9218**	0.5514
Set Workshop Scene	28.0	55.0	27.0	**32.39**	20.04	**0.9473**	0.5858

Table 5.6: Numerical results on the MPI dataset [6]

$\alpha = 16$	Disparity(pix)			PSNR(dB)		SSIM	
	d_{min}	d_{max}	d_{range}	CycleST	CycleLF	CycleST	CycleLF
average				32.10	**35.87**	0.9782	**0.9899**
Bikes	-14.0	9.5	23.5	30.64	**33.64**	0.9695	**0.9857**
Workshop	-6.5	16.5	23.0	33.55	**38.09**	0.9869	**0.9940**

MPI and CIVIT dataset are shown in Table 5.5, Table 5.6 and Table 5.7, respectively. As we can observe, our method CycleLF performs worse remove on the HDCA as this dataset has wider disparity range ($25 \sim 55$ pixels). However, CycleLF shows the advantage of video interpolation based methods on the MPI and CIVIT datasets as they has more narrow disparity range (< 16.5 pixels). Please note this experiment did not use any light field from these three datasets in the retraining step, however it is reasonable to expect improvement if the re-training dataset would include such light fields, in particular wide baseline light field from the HDCA dataset.

5.4 Conclusion

In this chapter, we proposed a novel self-supervised framework to reconstruct dense light fields by synthesizing novel intermediate light field views. To adopt small-sized light field datasets, we introduced the cycle consistency mechanism to fine-tune a pre-trained video

Table 5.7: Numerical results on the CIVIT dataset [7]

$\alpha = 16$	Disparity(pix)			PSNR(dB)		SSIM	
	d_{min}	d_{max}	d_{range}	CycleST	CycleLF	CycleST	CycleLF
average				35.91	**38.25**	0.9767	**0.9836**
Castle	-2.0	12.0	14.0	34.79	**36.54**	0.9606	**0.9659**
Dragon	-9.0	7.0	16.0	39.52	**40.95**	0.9875	**0.9880**
Flowers	-6.5	7.5	14.0	34.19	**37.66**	0.9841	**0.9909**
Holiday	-8.0	6.0	14.0	30.78	**34.04**	0.9628	**0.9800**

interpolation method in a self-supervised fashion. In this context, this method does not require paired ground-truth and is able to use for any low angular resolution light field input. The proposed method outperforms other approaches on various light fields, in particular, given handling wide disparity inputs. In addition, our method can be adapted to any video interpolation approach, and let any 2D video interpolation into applying to light field data. For future work, we may focus on adopting the proposed method to more challenging scenarios, such as very sparse light fields captured by camera arrays. This may require additional priors to handle the sparsity.

6 Conclusions and Outlook

Light field reconstruction is crucial for practical applications and still poses many challenges as an open research problem. In this work we investigate efficient and accurate light field reconstruction from various perspectives, including depth based rendering, deep learning based view synthesis and view selection strategies. In this chapter, we first summarize the main results and contributions, then we also outline potential future research.

6.1 Conclusions

One possible solution to reconstruct dense light fields is a cascade of geometry estimation in the spatio-angular domain and the forward warping of input views to generate novel views. To accomplish the reconstruction, extracting accurate depth using comprehensive high-dimensional information captured by the light field becomes the core component. In Chapter 3, we adapt optical flow based disparity estimation to the angular dimension of the light field. Our framework combines sparse initialization with edge-aware filtering, which preserves local geometric details and shows a reduced computational expense compared to global optimization. Inherent consistency is reinforced by employing edge-aware filtering in the angular domain. This approach demonstrates efficiency and accuracy compared to state-of-the-art depth estimation methods from light fields. Furthermore, compared to deep learning based methods, it doesn't require huge amounts of data as prior knowledge.

Light fields provide abundant high dimensional information, which causes significant computation complexity to process all views. View selection strategies plays a crucial role in determining the most important views and maximizing the performance of light field processing. However, even though light field reconstruction has been extensively studied in many previous works, existing approaches don't pay enough attention on carefully selecting views from a 4D light field. In Chapter 4, we highlight the importance of view selection strategies by comparing the influence of different strategies on the performance of light field processing. The evaluation of this study is performed using one selected benchmark method, which is utilized to compare three strategies for subsampling and six for reconstruction following the same principle. Our experimental results reveal the advantages of specific strategies which would be beneficial for further applications.

While application of deep learning in image processing is booming, it is promising to utilize the approximation capability of data-driven methods to reconstruct high dimensional information of a light field. However, compared to the size of general single image datasets, light field datasets are generally small sized due to the high expense of corresponding acquisition. In Chapter 5, we adapt the concept of cycle consistency to enable the training in a self-supervised fashion, which reduces the need for training data compared to conventional learning methods. We also adapt a state-of-the-art video interpolation approach to the angular domain, which is capable of transferring a comprehensive model extracted from large-scale video datasets to the light field domain. Experiments demonstrate the success in visually producing promising results and numerically outperforming state-of-the-art methods on both synthetic and real-world datasets.

6.2 Future Work

This thesis introduces several novel ideas about generalized reconstruction of light fields. In this final section, we will outline some interesting directions for refinement or application that could be worth to investigate further.

In Chapter 3, we incorporate an optical flow based method to estimate geometry information from light fields. This shows success in depth estimation and view synthesis for light fields. It mainly is dependent to the estimation on pixel intensity and it would be interesting to involve perceptual information as prior knowledge, such as contextual awareness [152]. Such an improved depth estimator would be beneficial for the consecutive rendering application. Moreover, the depth estimation step can be integrated to a CNN and more high level features can be extracted by various deep learning techniques. This could be helpful for handling scenes for instance with lighting changes or large displacements, which is challenging for conventional methods.

In Chapter 3 and Chapter 5, we demonstrate light field view synthesis solutions along one angular dimension. Our method demonstrates advantages among state-of-the-art methods, however, there still is space for improvement. For future attention, exploiting inherent consistency of the 4D structure of light fields may be meaningful to unleash the potential power hidden in light fields. For example, one potential direction could be to utilize 3D/4D convolutions or novel loss functions to extract and utilize features in higher dimensions.

Currently, we focus on light fields with limited range of disparity, which are usually captured by acquisition systems with narrow baseline between two adjacent lenses. There are still many other issues while light fields are employed in practical applications, such as very sparse light fields [28], unstructured light fields [107] and more. Large displacement could be the main issue in these two mentioned cases, and we believe one potential solution for this problem would be maintaining the identification of pixels or objects along the angular dimension, which could be implemented with deep learning techniques such as recognition, tracking or attention mechanisms.

In Chapter 4, we present the significant influence of the view selection strategy on light field view processing. We utilize the simplest evaluation to select the optimal view selection strategy. We believe there are more advanced techniques that can be employed to analyse the "redundancy" of light field views. One attempt would be integrating attention mechanisms to determine views that contribute most [90] and optimizing the different strategies utilizing views from specific locations regarding different tasks.

Bibliography

[1] M. Rerabek and T. Ebrahimi. New light field image dataset. In *Proceedings of the International Conference on Quality of Multimedia Experience*, 2016.

[2] Inria Lytro Illum dataset. `http://www.irisa.fr/temics/demos/lightField/CLIM/DataSoftware.html`. accessed: 26-01-2018.

[3] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*. Springer, 2016.

[4] V. vaish and a. adams. the (new) stanford light field archive. `http://lightfield.stanford.edu`. accessed: 01-09-2020.

[5] Matthias Ziegler, Ron op het Veld, Joachim Keinert, and Frederik Zilly. Acquisition system for dense lightfield of large scenes. In *2017 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2017.

[6] Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafał Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Towards a quality metric for dense light fields. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Sergio Moreschini, Filipe Gama, Robert Bregovic, and Atanas Gotchev. Civit dataset: Horizontal-parallax-only densely-sampled light-fields.

[8] Marc Levoy. Light fields and computational imaging. *Computer*, 39(8):46–55, 2006.

[9] Y. Chen, M. Alain, and A. Smolic. Fast and accurate optical flow based depth map estimation from light fields. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, 2017.

[10] Y. Chen, M. Alain, and A. Smolic. A study of efficient light field subsampling and reconstruction strategies. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, 2020.

[11] Yang Chen, Martin Alain, and Aljosa Smolic. Self-supervised light field view synthesis using cycle consistency. In *2020 IEEE 22st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020.

[12] Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. Light field microscopy. In *ACM SIGGRAPH 2006 Papers*, pages 924–934. 2006.

[13] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light Field Photography with a Hand-Held Plenoptic Camera. Technical report, Stanford University CSTR, Apr. 2005.

[14] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 31–42, 1996.

[15] Fu-Chung Huang, David P Luebke, and Gordon Wetzstein. The light field stereoscope. In *SIGGRAPH Emerging Technologies*, pages 24–1, 2015.

[16] Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, and Ramesh Raskar. Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting. 2012.

[17] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017.

[18] Todor georgiev. 100 years light-field. http://www.tgeorgiev.net/Lippmann/index.html.

[19] Edward H Adelson, James R Bergen, et al. *The plenoptic function and the elements of early vision*, volume 2. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of . . . , 1991.

[20] Leonard McMillan and Gary Bishop. Plenoptic modeling: An image-based rendering system. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 39–46. ACM, 1995.

[21] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996.

[22] Anders Heyden and Marc Pollefeys. Multiple view geometry. *Emerging topics in computer vision*, 3:45–108, 2005.

[23] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International journal of computer vision*, 1(1):7–55, 1987.

[24] Insung Ihm, Sanghoon Park, and Rae Kyoung Lee. Rendering of spherical light fields. In *Proceedings The Fifth Pacific Conference on Computer Graphics and Applications*, pages 59–68. IEEE, 1997.

[25] Emilio Camahort, Apostolos Lerios, and Donald Fussell. Uniformly sampled light fields. In *Eurographics Workshop on Rendering Techniques*, pages 117–130. Springer, 1998.

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

[27] Neus Sabater, Guillaume Boisson, Benoit Vandame, Paul Kerbiriou, Frederic Babon, Matthieu Hog, Remy Gendrot, Tristan Langlois, Olivier Bureller, Arno Schubert, et al. Dataset and pipeline for multi-view lightfield video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1743–1753. IEEE, 2017.

[28] Thorsten Herfet, Tobias Lange, and Harini Priyadarshini Hariharan. Enabling multiview-and light field-video for veridical visual experiences. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 1705–1709. IEEE, 2018.

[29] Jonas Trottnow, Simon Spielmann, Tobias Lange, Kelvin Chelli, Marek Solony, Pavel Smrz, Pavel Zemcik, Weston Aenchbacher, Mairéad Grogan, Martin Alain, et al. The potential of light fields in media productions. In *SIGGRAPH Asia 2019 Technical Briefs*, pages 71–74. 2019.

[30] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008. accessed: 01-09-2020.

[31] Densely sampled light field datasets. https://civit.fi/icme-2020-grand-challenge-on-densely-sampled-light-field-reconstruction/. accessed: 01-11-2020.

[32] The raytrix technology. https://raytrix.de/. accessed: 01-10-2020.

[33] The lytro illum camera. https://www.lytro.com/imaging. accessed: 01-10-2020.

[34] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1027–1034, 2013.

[35] Pierre Matysiak, Mairéead Grogan, Mikaël Le Pendu, Martin Alain, Emin Zerman, and Aljosa Smolic. High quality light field extraction and post-processing for raw plenoptic data. *IEEE Transactions on Image Processing*, 29:4188–4203, 2020.

[36] Alkhazur Manakov, John Restrepo, Oliver Klehm, Ramon Hegedus, Elmar Eisemann, Hans-Peter Seidel, and Ivo Ihrke. A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging. 2013.

[37] Ryan S Overbeck, Daniel Erickson, Daniel Evangelakos, and Paul Debevec. The making of welcome to light fields vr. In *ACM SIGGRAPH 2018 Talks*, pages 1–2. 2018.

[38] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020.

[39] Michael Broxton, Logan Grosenick, Samuel Yang, Noy Cohen, Aaron Andalman, Karl Deisseroth, and Marc Levoy. Wave optics theory and 3-d deconvolution for the light field microscope. *Optics express*, 21(21):25418–25439, 2013.

[40] Xing Lin, Jiamin Wu, Guoan Zheng, and Qionghai Dai. Camera array based light field microscopy. *Biomedical optics express*, 6(9):3179–3189, 2015.

[41] Robert Prevedel, Young-Gyu Yoon, Maximilian Hoffmann, Nikita Pak, Gordon Wetzstein, Saul Kato, Tina Schrödel, Ramesh Raskar, Manuel Zimmer, Edward S Boyden, et al. Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy. *Nature methods*, 11(7):727–730, 2014.

[42] Elliott Kwan, Yi Qin, and Hong Hua. Development of a light field laparoscope for depth reconstruction. In *3D Image Acquisition and Display: Technology, Perception and Applications*, pages DW1F–2. Optical Society of America, 2017.

[43] A Orth, M Ploschner, ER Wilson, IS Maksymov, and BC Gibson. Optical fiber bundles: Ultra-slim light field imaging probes. *Science advances*, 5(4):eaav1555, 2019.

[44] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4474, 2015.

[45] Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. Synthetic defocus and look-ahead autofocus for casual videography. *ACM Transactions on Graphics (TOG)*, 38(4):1–16, 2019.

[46] Jinsun Park, Yu-Wing Tai, Donghyeon Cho, and In So Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1736–1745, 2017.

[47] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M Seitz. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2015.

[48] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: shallow depth of field from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018.

[49] Aaron Isaksen, Leonard McMillan, and Steven J Gortler. Dynamically reparameterized light fields. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 297–306, 2000.

[50] Marc Levoy, Billy Chen, Vaibhav Vaish, Mark Horowitz, Ian McDowall, and Mark Bolas. Synthetic aperture confocal imaging. *ACM Transactions on Graphics (ToG)*, 23(3):825–834, 2004.

[51] Ren Ng. Fourier slice photography. In *ACM SIGGRAPH 2005 Papers*, pages 735–744. 2005.

[52] Vaibhav Vaish, Gaurav Garg, Eino-Ville Talvala, Emilio Antunez, Bennett Wilburn, Mark Horowitz, and Marc Levoy. Synthetic aperture focusing using a shear-warp factorization of the viewing transform. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 129–129. IEEE, 2005.

[53] Martin Alain, Weston Aenchbacher, and Aljosa Smolic. Interactive light field tilt-shift refocus with generalized shift-and-sum. In *European Light Field Imaging Workshop (ELFI)*, 2019.

[54] Shachar Ben Dayan, David Mendlovic, and Raja Giryes. Deep sparse light field refocusing. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual*

Event, UK, September 7-10, 2020. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0160.pdf.

[55] Steven Maesen, Patrik Goorts, and Philippe Bekaert. Omnidirectional free viewpoint video using panoramic light fields. In *2016 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2016.

[56] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

[57] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *European Conference on Computer Vision*, pages 441–459. Springer, 2020.

[58] Matthias Ziegler, Andreas Engelhardt, Stefan Müller, Joachim Keinert, Frederik Zilly, Siegfried Foessel, and Katja Schmid. Multi-camera system for depth based visual effects and compositing. In *Proceedings of the 12th European Conference on Visual Media Production*, pages 1–10, 2015.

[59] Creal30 "see real 3d". https://www.creal.com/. accessed: 01-11-2020.

[60] Avegant light field technology. https://www.avegant.com/. accessed: 01-11-2020.

[61] David Fattal, Zhen Peng, Tho Tran, Sonny Vo, Marco Fiorentino, Jim Brug, and Raymond G Beausoleil. A multi-directional backlight for a wide-angle, glasses-free three-dimensional display. *Nature*, 495(7441):348–351, 2013.

[62] Looking glass company. https://lookingglassfactory.com/. accessed: 01-11-2020.

[63] Matthew Hirsch, Daniel Leithinger, and Thomas Baran. Lumii: Diy light field prints. In *ACM SIGGRAPH 2016 Studio*, pages 1–2. 2016.

[64] Sony spatial reality display. https://www.sony.com/electronics/spatial-reality-display/elf-sr1. accessed: 01-11-2020.

[65] Hae Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu Wing Tai, and In So Kweon. Accurate depth map estimation from a lenslet light field camera. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1547–1555, 2015.

[66] Neus Sabater, Mozhdeh Seifi, Valter Drazic, Gustavo Sandri, and Patrick Pérez. Accurate Disparity Estimation for Plenoptic Images. volume 8926 of *Lecture Notes in Computer Science*, pages 548–560. Springer International Publishing, 2015.

[67] How Accurate Can Block Matches Be in Stereo Vision? *SIAM Journal on Imaging Sciences*, 4(1):472–500, 2011.

[68] Julia Navarro and Antoni Buades. Robust and Dense Depth Estimation for Light Field Images. *IEEE Transactions on Image Processing*, 26(4):1873–1886, apr 2017.

[69] Alessandro Neri, Marco Carli, and Federica Battisti. A multi-resolution approach to depth field estimation in dense image arrays. *Proceedings - International Conference on Image Processing, ICIP*, 2015-Decem:3358–3362, 2015.

[70] Stefan Heber and Thomas Pock. Shape from Light Field Meets Robust PCA. Number 836630, pages 751–767. 2014.

[71] Can Chen, Haiting Lin, Zhan Yu, Sing Bing Kang, and Jingyi Yu. Light field stereo matching using bilateral statistics of surface cameras. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1518–1525, 2014.

[72] Ting-Chun Wang, Alexei A. Efros, and Ravi Ramamoorthi. Occlusion-Aware Depth Estimation Using Light-Field Cameras. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3487–3495, 2015.

[73] Lipeng Si and Qing Wang. *Dense Depth-Map Estimation and Geometry Inference from Light Fields via Global Optimization*, pages 83–98. Springer International Publishing, 2017.

[74] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73:1–73:12, July 2013.

[75] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. *Proceedings of the IEEE International Conference on Computer Vision*, 2:673–680, 2013.

[76] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4D light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, jun 2012.

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

[78] Ole Johannsen, Antonin Sulc, and Bastian Goldluecke. What Sparse Light Field Coding Reveals about Scene Structure. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3262–3270. IEEE, jun 2016.

[79] Ting-Chun Wang, Jun-Yan Zhu, Ebi Hiroaki, Manmohan Chandraker, Alexei A Efros, and Ravi Ramamoorthi. A 4d light-field dataset and cnn architectures for material recognition. In *European Conference on Computer Vision*, pages 121–138. Springer, 2016.

[80] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):193, 2016.

[81] Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field reconstruction using deep convolutional network on epi. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017, page 2, 2017.

[82] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 24–32, 2015.

[83] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3746–3754, 2016.

[84] Stefan Heber, Wei Yu, and Thomas Pock. U-shaped networks for shape from light field. In *BMVC*, volume 3, page 5, 2016.

[85] Stefan Heber, Wei Yu, and Thomas Pock. Neural epi-volume networks for shape from light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2252–2260, 2017.

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

[87] Wenhui Zhou, Linkai Liang, Hua Zhang, Andrew Lumsdaine, and Lili Lin. Scale and orientation aware epi-patch learning for light field depth estimation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2362–2367. IEEE, 2018.

[88] Haoxin Ma, Haotian Li, Zhiwen Qian, Shengxian Shi, and Tingting Mu. Vommanet: an end-to-end network for disparity estimation from reflective and textureless light field images. *arXiv preprint arXiv:1811.07124*, 2018.

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

[90] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *AAAI*, pages 12095–12103, 2020.

[91] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017.

[92] Zhouchen Lin and Heung-Yeung Shum. A geometric analysis of light field rendering. *International Journal of Computer Vision*, 58(2):121–138, 2004.

[93] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 307–318, 2000.

[94] Cha Zhang and Tsuhan Chen. Spectral analysis for sampling image-based rendering data. *IEEE Transactions on Circuits and Systems for Video Technology*, 13 (11):1038–1050, 2003.

[95] Minh N Do, Davy Marchand-Maillet, and Martin Vetterli. On the bandwidth of the plenoptic function. *IEEE Transactions on Image Processing*, 21(2):708–717, 2011.

[96] Christopher Gilliam, Pier-Luigi Dragotti, and Mike Brookes. On the spectrum of the plenoptic function. *IEEE transactions on image processing*, 23(2):502–516, 2013.

[97] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017.

[98] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019.

[99] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2013.

[100] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2243–2251, 2017.

[101] Jing Jin, Junhui Hou, Hui Yuan, and Sam Kwong. Learning light field angular super-resolution via a geometry-aware network. In *AAAI*, pages 11141–11148, 2020.

[102] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[103] Henry Wing Fung Yeung, Junhui Hou, Jie Chen, Yuk Ying Chung, and Xiaoming Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 137–152, 2018.

[104] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5): 2319–2330, 2019.

[105] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. Learning fused pixel and feature-based view reconstructions for light fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2555–2564, 2020.

[106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[107] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion:

Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[108] Yunlong Wang, Fei Liu, Zilei Wang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. End-to-end view synthesis for light field imaging with pseudo 4DCNN. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–348, 2018.

[109] Gaochang Wu, Yebin Liu, Lu Fang, Qionghai Dai, and Tianyou Chai. Light field reconstruction using convolutional network on epi and extended applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1681–1694, 2018.

[110] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[111] Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field reconstruction using deep convolutional network on EPI. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6319–6327, 2017.

[112] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Advances in neural information processing systems*, pages 1033–1041, 2009.

[113] Gaochang Wu, Yebin Liu, Qionghai Dai, and Tianyou Chai. Learning sheared epi structure for light field reconstruction. *IEEE Transactions on Image Processing*, 28(7):3261–3273, 2019.

[114] Lixin Shi, Haitham Hassanieh, Abe Davis, Dina Katabi, and Fredo Durand. Light field reconstruction using sparsity in the continuous fourier domain. *ACM Transactions on Graphics (TOG)*, 34(1):1–13, 2014.

[115] Suren Vagharshakyan, Robert Bregovic, and Atanas Gotchev. Light field reconstruction using shearlet transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):133–147, 2017.

[116] Suren Vagharshakyan, Robert Bregovic, and Atanas Gotchev. Accelerated shearlet-domain light field reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1082–1091, 2017.

[117] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[118] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, Cen Rao, and Michael Isnardi. Bilateral filtering-based optical flow estimation with occlusion detection. In *European conference on computer vision*, pages 211–224. Springer, 2006.

[119] Manuel Lang, Oliver Wang, Tunc Aydin, Aljoscha Smolic, and Markus Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics (ToG)*, 31(4):34, 2012.

[120] Wenbin Li, Darren Cosker, Matthew Brown, and Rui Tang. Optical flow estimation using laplacian mesh energy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2435–2442, 2013.

[121] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 28(3):24, 2009.

[122] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172, 2015.

[123] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016.

[124] David W Murray and Bernard F Buxton. Scene segmentation from visual motion using global optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):220–228, 1987.

[125] Sebastian Volz, Andres Bruhn, Levi Valgaerts, and Henning Zimmer. Modeling temporal coherence for optical flow. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1116–1123. IEEE, 2011.

[126] Matthias Hoeffken, Daniel Oberhoff, and Marina Kolesnik. Temporal prediction and spatial regularization in differential optical flow. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 576–585. Springer, 2011.

[127] Michael Schaffner, Florian Scheidegger, Lukas Cavigelli, Hubert Kaeslin, Luca Benini, and Aljosa Smolic. Towards edge-aware spatio-temporal filtering in real-time. *IEEE Transactions on Image Processing*, 27(1):265–280, 2018.

[128] Danny Barash. Fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):844–847, 2002.

[129] Ron Kimmel, Nir Sochen, and Ravi Malladi. From high energy physics to low level vision. In *International Conference on Scale-Space Theories in Computer Vision*, pages 236–247. Springer, 1997.

[130] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998.

[131] Tunç Ozan Aydin, Nikolce Stefanoski, Simone Croci, Markus Gross, and Aljoscha Smolic. Temporally coherent local tone mapping of hdr video. *ACM Transactions on Graphics (TOG)*, 33(6):196, 2014.

[132] Eduardo SL Gastal and Manuel M Oliveira. Domain transform for edge-aware image and video processing. In *ACM Transactions on Graphics (ToG)*, volume 30, page 69. ACM, 2011.

[133] Zeev Farbman, Raanan Fattal, Dani Lischinski, and Richard Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. In *ACM Transactions on Graphics (TOG)*, volume 27, page 67. ACM, 2008.

[134] Raanan Fattal. Edge-avoiding wavelets and their applications. *ACM Transactions on Graphics (TOG)*, 28(3):22, 2009.

[135] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *European conference on computer vision*, pages 1–14. Springer, 2010.

[136] Eduardo SL Gastal and Manuel M Oliveira. High-order recursive filtering of non-uniformly sampled signals for image and video processing. In *Computer Graphics Forum*, volume 34, pages 81–93. Wiley Online Library, 2015.

[137] Sylvain Paris, Samuel W Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Trans. Graph.*, 30(4): 68–1, 2011.

[138] Mathieu Aubry, Sylvain Paris, Samuel W Hasinoff, Jan Kautz, and Frédo Durand. Fast local laplacian filters: Theory and applications. *ACM Transactions on Graphics (TOG)*, 33(5):167, 2014.

[139] Antonio Criminisi, Toby Sharp, Carsten Rother, and Patrick Pérez. Geodesic image and video editing. *ACM Trans. Graph.*, 29(5):134–1, 2010.

[140] Cevahir Cigla and A Aydın Alatan. Information permeability for stereo matching. *Signal Processing: Image Communication*, 28(9):1072–1088, 2013.

[141] Peyman Milanfar. A tour of modern image filtering: New insights and methods, both practical and theoretical. *IEEE Signal Processing Magazine*, 30(1):106–128, 2013.

[142] Martin Alain and Aljosa Smolic. A spatio-angular binary descriptor for fast light field inter view matching. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2636–2640. IEEE, 2020.

[143] David Matheson Young and Werner Rheinboldt. Iterative solution of large linear systems. 1971.

[144] Li Xu, Jiaya Jia, and Yasuyuki Matsushita. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (9):1744–1757, 2012.

[145] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William Freeman. Sift flow: Dense correspondence across different scenes. *Computer vision–ECCV 2008*, pages 28–42, 2008.

[146] Linchao Bao, Qingxiong Yang, and Hailin Jin. Fast edge-preserving patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3534–3541, 2014.

[147] Joan Sol Roo and Christian Richardt. Temporally coherent video de-anaglyph. In *ACM SIGGRAPH 2014 Posters*, page 75. ACM, 2014.

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

[149] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017.

[150] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017.

[151] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[152] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018.

[153] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.

[154] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 2019.

[155] Yuan Gao and Reinhard Koch. Parallax view generation for static scenes using parallax-interpolation adaptive separable convolution. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–4. IEEE, 2018.

[156] Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field reconstruction using deep convolutional network on EPI. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6319–6327, 2017.

[157] Anna Alperovich, Ole Johannsen, Michael Strecke, and Bastian Goldluecke. Light field intrinsics with a deep encoder-decoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9145–9154, 2018.

[158] Fan Wang, Qixing Huang, and Leonidas J Guibas. Image co-segmentation via consistent functional maps. In *Proceedings of the IEEE international conference on computer vision*, pages 849–856, 2013.

[159] Fan Wang, Qixing Huang, Maks Ovsjanikov, and Leonidas J Guibas. Unsupervised multi-class joint image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3142–3149, 2014.

[160] Xiaowei Zhou, Menglong Zhu, and Kostas Daniilidis. Multi-image matching via fast alternating minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4032–4040, 2015.

[161] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2015.

[162] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016.

[163] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[164] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[165] Kyle Wilson and Noah Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 513–520, 2013.

[166] Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1426–1433. IEEE, 2010.

[167] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[168] H Russell Bernard. *Research methods in anthropology: Qualitative and quantitative approaches*. Rowman & Littlefield, 2017.

[169] Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8794–8802, 2019.

[170] Fitsum A Reda, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J Shih, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Unsupervised video interpolation using cycle consistency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 892–900, 2019.

[171] Yuan Gao, Robert Bregovic, and Atanas P. Gotchev. Self-supervised light field reconstruction using shearlet transform and cycle consistency. *IEEE Signal Process. Lett.*, 27:1425–1429, 2020. doi: 10.1109/LSP.2020.3008082. URL https://doi.org/10.1109/LSP.2020.3008082.

[172] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[173] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[174] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[175] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

[176] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[177] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[178] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[179] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[180] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.

[181] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

[182] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[183] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[184] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[185] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2017.

[186] Pierre Matysiak, Mairéad Grogan, Mikaël Le Pendu, Martin Alain, and Aljosa Smolic. A pipeline for lenslet light field quality enhancement. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 639–643. IEEE, 2018.

[187] Yuan Gao, Robert Bregovic, Reinhard Koch, and Atanas Gotchev. Drst: Deep residual shearlet transform for densely sampled light field reconstruction. *arXiv preprint arXiv:2003.08865*, 2020.