Single-Shot View Synthesis using a Multiplexed Light Field Camera

Shamus Li

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-192
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-192.html

November 13, 2024

Single-Shot View Synthesis using a Multiplexed Light Field Camera

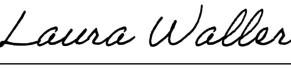
by Shamus Li

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee



Professor Laura Waller
Research Advisor

11/13/24

(Date)

⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆

Professor Ren Ng
Second Reader

11/13/2024

(Date)

Single-Shot View Synthesis using a Multiplexed Light Field Camera

by

Shamus Li

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Laura Waller, Chair
Professor Ren Ng

Spring 2024

Single-Shot View Synthesis using a Multiplexed Light Field Camera

Abstract

Single-Shot View Synthesis using a Multiplexed Light Field Camera

by

Shamus Li

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Laura Waller, Chair

Recent advancements in imaging technologies have shifted from traditional 2D image capture to more sophisticated methods that aim to capture additional dimensions—spatial, temporal, etc.—of a given scene. We present an approach to single-shot view synthesis using a multiplexed light field camera, where sub-images are designed to overlap with each other to achieve higher spatial and temporal resolution compared to conventional light field imaging. We use a single capture from our optical system to achieve novel view synthesis.

Our system captures light fields through a lens array that intentionally overlaps views, enhancing both resolution and depth of field. This multiplexing approach is complemented by a calibration process that aligns virtual camera poses, facilitating accurate reconstruction without repeated pose estimation. We modify the forward model of Gaussian Splatting to implicitly represent and reconstruct the light field from the multiplexed measurements.

We present synthetic experimental results that demonstrate the efficacy of our system in generating wide-angle, photorealistic 3D reconstructions of small scenes both in simulation and the real world, and discuss extensions to a physical system. We achieve an optical field of view of more than 70 degrees, and are able to accurately reconstruct more than 120 degrees with a single shot. Our physical system achieves 1.9 rays/pixel of multiplexing, a 90% increase in pixel information over a light field imaging system with no overlapping, and we demonstrate higher-quality reconstructions on synthetic scenes with up to 2.5 rays/pixel of multiplexing when compared to both traditional light field images as well as monocular Gaussian Splatting. Our method represents a potential step forward in the practical application of view synthesis, particularly in dynamic environments with few cameras.

To my family.

Contents

List of Figures

Acknowledgments

I would first like to thank Kristina Monakhova and Kyrollos Yanny for helping a curious freshman discover the field of computational imaging for the first time. It was those weekly meetings while I was stuck in my room that helped me decide that I wanted to pursue an advanced degree. I would like to thank the whole of WallerLab for sharing lively discussions with me and offering me your wisdom about everything from research to the outdoors. This work wouldn't have been possible without support from Sara Fridovich-Keil, Ruiming Cao, and Kevin Zhou, whose expertise was instrumental for achieving my research goals. Whenever I felt lost, asking them has often been the right answer. In addition, some of the work presented here was done jointly with Vi Tran, who is a fantastic person to work alongside, and the rigour of their research is much appreciated. I would like to thank Professor Ren Ng for his feedback on this report and for being an inspiration. Lastly, I would like to thank my advisor Professor Laura Waller for her guidance both in shaping my experiments and in navigating a career in academia. I am extremely grateful to have had such great mentorship throughout my time at Berkeley.

Chapter 1

Introduction

The evolution of imaging technologies, from traditional film-based cameras to modern digital sensors, have brought about significant advances in how we capture and interpret the world around us. Conventionally, cameras have been designed to capture two-dimensional images, focusing on the production of sharp, well-exposed photographs that represent a single perspective of a scene. However, the dimensionality of light extends far beyond the confines of 2D image planes. Light interacting with the environment carries information not only about intensity, but also about direction, wavelength, and time. A parameterization of this is the plenoptic function—$P(\theta, \phi, \lambda, t, V_x, V_y, V_z)$, where θ and ϕ is the direction of light, λ is the wavelength, t is time, and V_x, V_y, V_z is 3D origin of the light ray—which represents every possible image from every viewpoint in a particular space-time chunk [2]. It is therefore necessary to map this higher-dimension to a 2D grid to capture this lost information, leading to sacrifices either in spatial or temporal resolution. The primary purpose of this work is to design an optical coding to limit these tradeoffs as much as possible.

The focus of my work is on rendering images from more viewpoints than were actually captured, a technique called novel view synthesis. This is achieved by not only capturing the 2D intensity of light that hits each pixel, but also measuring the amount of light travelling along each ray that intersects the sensor. We can model this ray in 5D by removing time and wavelength from the plenoptic function, or in 4D as a parameterization of a line that intersects two planes [15].

Traditionally, this representation, known as a light field, was explicitly represented and required a dense grid of views to be captured. Recent techniques such as Neural Radiance Fields (NeRF) have revolutionized the field by learning implicit scene representations that enable high-quality image synthesis from novel viewpoints [14]. NeRF and its derivatives can reconstruct a 3D scene from a relatively sparse set of input images captured from different viewpoints. However, the capture of these views typically takes a long time and assumes a static scene, heavily limiting their applicability in dynamic environments that change over time.

Light field cameras, which simultaneously record multiple perspectives in one sensor measurement, offer a potential solution to this problem. By capturing both spatial and

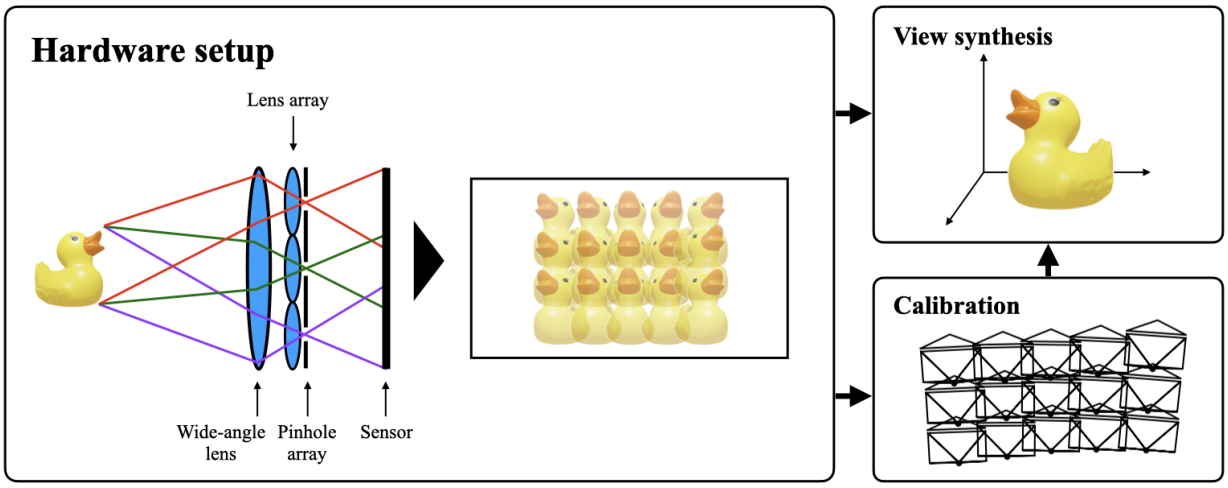


Figure 1.1: We present an imaging system for single-shot view synthesis using a multiplexed light field camera. The captured image on the sensor consists of multiple overlapping views. The captured data is then processed through our view synthesis pipeline to generate novel views of the scene. The system is calibrated by capturing images through individual lenslets, allowing estimation of camera poses using structure-from-motion techniques.

angular information of light rays, light field cameras enable post-capture refocusing, depth estimation, and view synthesis. However, traditional light field cameras—from camera arrays to plenoptic cameras—face a fundamental trade-off between spatial resolution and angular resolution. Capturing more angular information typically results in a decrease in spatial resolution and vice versa.

This work introduces a novel approach to single-shot view synthesis using a multiplexed light field camera. By intentionally overlapping the views captured by a lens array, it is possible to achieve a higher space-bandwidth product than would be possible with non-overlapping monocular views. This is ideal for highly dynamic scenes in the mesoscale, making the system limited only by the capabilities of the camera sensor. In addition, by fixing the optics, we only need to calibrate the camera parameters once per camera, skipping a costly and potentially inaccurate pose estimation step in future reconstructions. We modify Gaussian Splatting to handle training from a single multiplexed image such that instead of rendering one image for each training pass, we render one image from each viewpoint in the camera and combine them to create the multiplexed image. We calibrate our camera using a traditional structure-from-motion pipeline. We demonstrate the efficacy of our system through both simulation and real-world experiments. We achieve an optical field of view of more than 70 degrees, and are able to accurately reconstruct more 120 degrees with a single shot. Our physical system achieves 1.9 rays/pixel of multiplexing, a 90% increase in pixel information over a light field imaging system with no overlapping, and we demonstrate

higher-quality reconstructions on synthetic scenes with up to 2.5 rays/pixel of multiplexing when compared to both traditional light field images as well as monocular Gaussian Splatting.

1.1 Related Work

Light Field Imaging

Light field cameras passively capture 4D space-angle information in a single shot, enabling 3D reconstructions, amongst other applications. Light field imaging has been a crucial research area in computational photography and computer vision, focusing on capturing the full dimensionality of light rays in a scene. The plenoptic function, introduced by Adelson and Bergen, parameterizes light rays by their position, direction, wavelength, and time, encapsulating the entirety of visual information available in a scene [2]. Unlike traditional imaging techniques that capture only the intensity of light at each point, light field imaging captures the intensity of light rays as a function of space and angle. This additional information enables computational capabilities not possible with conventional cameras.

The core component of a light field camera is a microlens array placed in front of the image sensor. Each microlens captures light rays from different directions and focuses them onto the sensor, allowing each pixel to receive light information from a specific direction. The captured light field data can be represented as a four-dimensional function, $L(u, v, s, t)$, where (u, v) denote spatial coordinates and (s, t) represent angular coordinates of the light rays. Light field cameras can be modeled as an array of cameras, each capturing a slightly different perspective of the scene. Consequently, the captured data comprises a series of sub-images, each representing a slightly different viewpoint. This multi-view data enables refocusing and depth of field changes, disparity and depth calculation, as well as 3D reconstruction [8].

Implementations of light field cameras, such as the plenoptic camera proposed by Adelson and Wang [1] and notably by Ng et al [15], use a microlens array placed in front of an image sensor to capture multiple views of a scene from slightly different perspectives in a single shot. An alternative light field camera design is a camera array, which allows for for new viewpoints to be generated by interpolating between captured images [27]. These works demonstrated the concept of interpreting 2D images as slices of a 4D light field function, facilitating efficient creation and display of new views without requiring depth information or feature matching.

However, traditional light field cameras face significant trade-offs between spatial and angular resolutions. Capturing more angular information typically results in a decrease in spatial resolution and vice versa. This trade-off limits the applicability of traditional light field cameras in scenarios requiring high-resolution imaging and wide fields of view. Subsequent works have aimed to improve the spatial and angular resolution trade-offs inherent in these systems. Georgiev and Intwala proposed a system using a hexagonal array of twenty larger lenslets in order to reduce gaps between lenslets [5]; Lumsdaine and Georgiev introduced the concept of the focused plenoptic camera, which improves the spatial resolution

by simply adjusting the placement of the microlens array relative to the sensor [10]; and Perwaß and Wietzke presented a 3D camera which achieves improved depth estimation with a multi-focal microlens array. While these methods improve upon traditional designs, they do not fully overcome the inherent trade-offs.

Lenslet array-based capture schemes have also been widely used in microscopy for 3D depth imaging [17, 22]. In particular, Fourier Light Field Microscopy (FLFM) has emerged as a powerful technique in computational microscopy. FLFM operates by placing a microlens array at the Fourier plane of the imaging system, which creates a three-dimensional shift-invariant point spread function (PSF), enabling the reconstruction of volumetric information from a single two-dimensional (2D) measurement. [6]. This approach has been further refined with techniques like Fourier Diffuserscope, which introduces a diffuser at the Fourier plane to encode additional spatial information and improve reconstruction quality [9]. Our work draws inspiration from FLFM but extends the concept to mesoscale imaging of objects in the millimeter to centimeter range.

The main idea of light field imaging is to encode additional angular information into the captured data, which can then enable synthetic refocusing, volume reconstruction, or neural reconstruction from a single sensor measurement. We introduce an optical system physically similar to that proposed by Georgiev and Intwala, but with a key difference: our system is designed to intentionally overlap the images from each lenslet onto the sensor, a new idea. By overlapping the views captured by the lens array, we effectively increase the amount of information—space-bandwith product—captured without sacrificing spatial resolution, enabling higher-resolution and wider field-of-view imaging.

Novel View Synthesis

Novel view synthesis requires recovery of a 3D representation of an object or scene from 2D input images. Existing methods often utilize point clouds [12], voxel grids [13], or signed distance functions [16] to represent the target. These approaches typically require a large set of training images and corresponding camera pose estimates to achieve accurate results. Practical applications of high-quality 3D reconstructions include generating 3D models for assets in animation, creating training environments for robotics simulations, and enhancing biological analysis.

Neural Radiance Fields (NeRFs) have emerged as a powerful technique for novel view synthesis [14]. NeRFs model appearance and geometry using radiance fields that map spatial coordinates and view direction to density and color values. This approach uses a dense set of images to train the network, which learns to predict the color and density of points in 3D space, allowing for high-quality view synthesis from novel viewpoints. Research has demonstrated that a small multi-layer perceptron (MLP) with positionally-encoded input coordinates can accurately represent a target scene [23]. Through standard volume rendering procedures, rays can be sampled, evaluated, and converted to image pixels, with the model optimizing the mean squared error between the outputted RGB values and the training images. The radiance field can be rendered as images, depth maps, or converted to a mesh

for downstream applications. NeRFs have demonstrated impressive results in capturing fine details and complex lighting effects, but they assume a stationary and unchanging target scene across all training images, rely on accurate camera pose estimates from structure-from-motion algorithms like COLMAP [20], and are slow and computationally expensive to train, taking hours for a single scene [14].

Significant optimizations have improved the efficiency of NeRF-based methods. For instance, techniques have dramatically increased training speed [24], and some approaches, such as Plenoxels, enable faster training without neural networks [19]. PixelNeRF and similar works suggest that training with a few input images might be feasible [28, 25]. However, these few-image input methods generally infer the missing views in the scene. Our system captures a larger area of the scene and encodes it into a single image, ensuring the training images more accurately represent the scene and allowing for real-time data capture.

Several extensions and improvements to NeRF have been proposed to address its limitations. D-NeRF adapts NeRF for dynamic scenes by incorporating temporal information, allowing for the synthesis of scenes that change over time [18]. MonoNeRF attempts to generalize NeRF to monocular videos, enabling view synthesis without precise camera poses [4]. However, these methods still face challenges in terms of training time and computational resources.

An alternative approach to view synthesis is Gaussian Splatting, which leverages the inherent sparsity in 3D scenes by representing scenes using 3D Gaussian functions–"Gaussians"–optimized for position, orientation, size, and color [7]. This method can render high-quality images in real time while preserving image reconstruction quality, making it a state-of-the-art technique for novel view synthesis.

We adapt Gaussian Splatting to handle multiplexed images captured by our light field camera. Because our images have a higher space-bandwidth product than traditional monocular views, we are able to create a higher-fidelity reconstruction than existing methods. The overarching goal is to achieve a wider field of view with our camera using intentionally multiplexed data, enabling efficient and accurate reconstruction of photorealistic volumes from a single capture without needing to predict or general additional views in the training data.

Compressed Sensing

Compressed sensing is an imaging technique that enables signals to be acquired with fewer measurements by exploiting the underlying structure of the signal for high-quality reconstruction [3]. Typically, capturing a signal requires measurements at twice the maximum spatial frequency of the signal—as per the Shannon-Nyquist sampling theorem—to ensure all information is captured. However, signals are often compressible, and the sum of more information can be captured with a single sensor pixel by spreading out the sparse information contained in the signal through multiplexing, effectively resulting in more useful information. One of the key benefits of compressed sensing is its ability to significantly reduce acquisition time and data storage requirements, which is particularly useful for high-speed or high-resolution 3D imaging applications.

Compressed sensing is particularly effective when signals exhibit sparsity in some domain. This is highly relevant to computational imaging applications, many of which aim to reconstruct a high-dimensional scene from a limited number of measurements. The compressed sensing paradigm represents a powerful tool in imaging system design, where the sensing hardware is viewed as an encoder rather than a direct signal approximator. This concept has already made a significant impact in fields such as MRI and computed tomography, accelerating scan speeds by reducing the number of samples required [11]. In compressed sensing, the sensing process involves capturing multiplexed measurements, which are linear combinations of the signal's components. These measurements are then processed using algorithms that exploit the sparsity of the signal to reconstruct the original high-dimensional data. This approach contrasts with traditional methods that directly sample each component of the signal individually. By encoding multiple dimensions of the optical image, compressed sensing enables the recovery of detailed scene information from fewer measurements. In the context of optical design, this raises the question of how to design optics that encode additional dimensions of optical images such that sparse recovery can successfully and accurately reconstruct the image. Specifically, in this work, we explore how optical design can be leveraged to extract larger space-bandwidth product light fields from a single measurement.

Our work employs compressed sensing in conjunction with multiplexed light field imaging to enhance the capabilities of traditional imaging systems. By integrating intentional overlapping views into the optical design, we can encode more scene information into each captured image. When compared to existing light field cameras, our approach achieves a higher space-bandwidth product with the same number of measurements. When compared to existing novel view synthesis techniques, our method significantly reduces the number of measurements needed for accurate reconstruction, improving the practicality of novel view synthesis, especially in dynamic environments.

Chapter 2

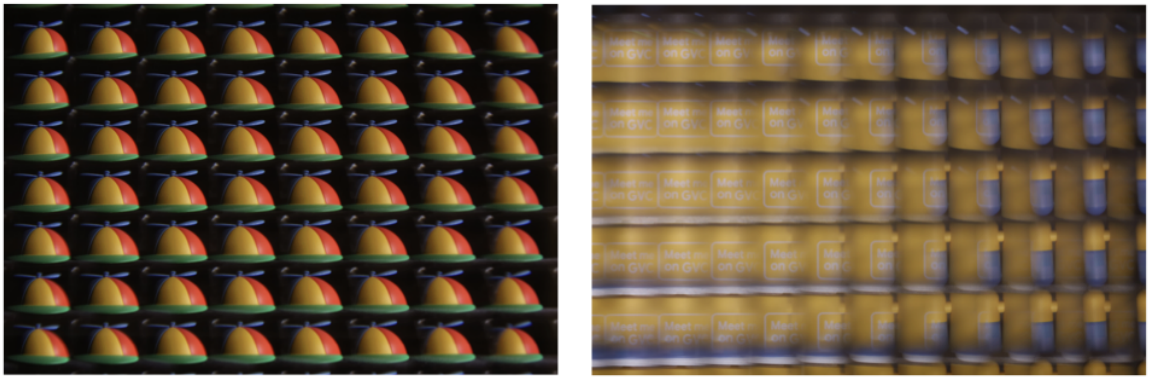
Building a Multiplexed Light Field Camera

The objective of this imaging system is to capture a comprehensive representation of a scene in a single shot by multiplexing multiple perspective images into a single image. The core of our optical design achieves this through a lens array that overlaps the views from each lenslet, providing an angular field of view and resolution not possible without overlapping. This approach is inspired by Fourier light field microscopy [6] but is applied here to mesoscale objects in the millimeter to centimeter range.

2.1 Optical Design

Our imaging system consists of a large-aperture photography lens placed backwards, a lens array, an aperture array, and a regular full-frame sensor (see Figure 2.3). The large-aperture main lens is reversed to enable focusing at closer distances, allowing us to image smaller objects placed very close to the lens (less than 2 cm). This proximity maximizes the angular range of captured light rays—the perspective shift between overlapping views. Although reversing the lens slightly increases magnification, it creates a larger perspective shift among the sub-lens images, maximizing the angular range of the overall light field capture.

The lens array, positioned immediately behind the main lens, consists of an array of small lenses (lenslets) that create separate overlapping views onto the sensor pixels. Our system aims to be a more affordable and compact alternative to a traditional camera array, such as proposed in [27], by capturing all sub-images onto a single sensor. In traditional imaging systems, any overlap would be clipped by the aperture so that each pixel on the sensor only receives light from one ray/sub-lens, but we propose controlling the amount of clipping to control the amount of overlap between the images from adjacent lenslets. By overlapping sub-lens views and reconstructing them computationally, we are able to capture more effective information with a wider baseline between the views. We use rays per pixel—where each ray comes from a different lenslet—as the overlap metric to balance the need for



No crosstalk Crosstalk

Figure 2.1: Example of captured images with optical crosstalk. Optical crosstalk occurs when light intended for one section of the sensor inadvertently reaches the area designated for another lens, causing undesired image artifacts.

perspective shift with the clarity of sub-images:

$$\text{Rays/pixel} = \frac{1}{N}\sum_{i=1}^{N} n_i \tag{2.1}$$

where N is the total number of pixels on the sensor, and n_i is the number of lenslets contributing light to pixel i. A value of 1 indicates no multiplexing, and higher values of the overlap metric correspond to increased multiplexing. A good rough heuristic is that adjacent images should overlap by approximately 50%, meaning that each sub-image shares half of its area on the sensor with neighboring sub-images, leaving the other half dedicated to that sub-image. This corresponds to of 1.5 rays per pixel.

In addition, the leftmost and rightmost perspectives of the scene, as well as the topmost and bottommost views, should have the largest possible baseline—the widest possible angular range. The ideal angular field of view (AFOV) of the system—the horizontal angle, from the optical axis, of light that can be captured by the lens—is dictated by the main lens' numerical aperture (NA) and focal length. Assuming an ideal thin lens and that the entire sensor is utilized, the AFOV can be calculated using the formula:

$$\text{AFOV} = 2\arctan\left(\frac{w}{2\cdot f}\right) \tag{2.2}$$

where w is the sensor width and f is the focal length of the main lens.

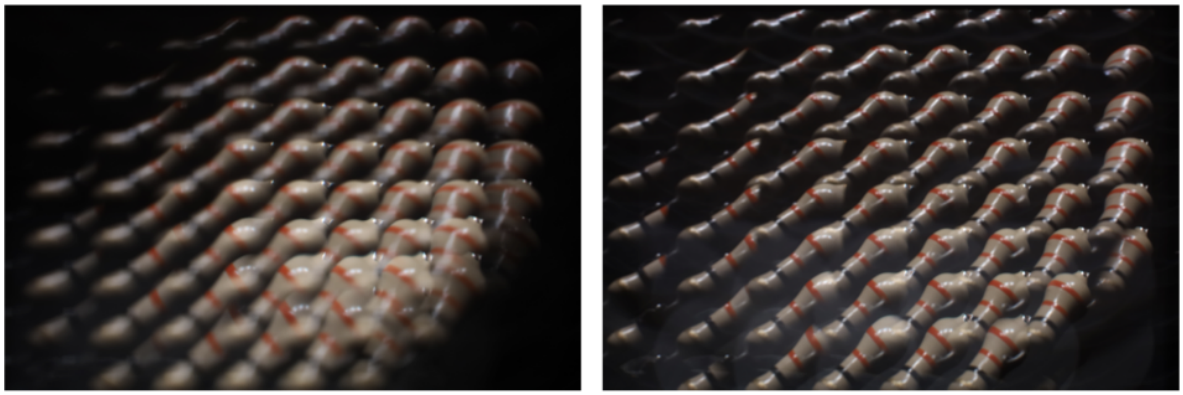


Figure 2.2: The aperture array mitigates optical crosstalk by blocking stray light between lenslets; increases the depth of field by limiting the effective aperture size for each sub-lens; and controls the amount of overlap. While it does block a significant amount of light, at the mesoscale this is not an issue.

An essential component in this design is the aperture array, a grid of 2 mm diameter small apertures placed at the center of each lenslet. The aperture array has three effects. First, it reduces optical crosstalk—when light rays intended for one lenslet overlap and spill over into another lenslet's view. The aperture array acts as a physical barrier to minimize stray light, ensuring each light ray is correctly mapped to its intended lenslet. Second, it also extends the effective depth-of-field by limiting the aperture size of each sub-image formed by the lenslets (see Figure 2.2. Lastly, it controls the amount of overlap between adjacent images; by using arrays with different aperture diameters, we can manage how much the sub-images overlap on the sensor. Figure 2.3 describes configuration of the optical system. All optical components are placed as close together as possible to reduce the effects of free-space propagation.

2.2 Methods

The choice of components is driven by the need to maximize the field of view. The main lens and lens array were selected to have similar cross-sectional diameters to ensure that the effects of the main lens are most effectively represented by the outer sub-lenses. For our system, we used a Canon TV-16 25mm $f/0.78$ lens for its extremely wide aperture, which allows a high amount of light gathering and angular coverage. The lens array, sourced

from Edmund Optics, measures 46 mm x 46 mm, with individual lenslets measuring 4 mm x 3 mm and an effective focal length of 38.10 mm. We use a subset of these lenslets in a 7 x 6 grid, resulting in 42 sub-views. The camera sensor used is a Canon RP full-frame camera sensor. We manufactured our own aperture array with 1 mm thick aluminum and an aperture diameter of 2 mm. Using our main lens with our sensor size, the calculated AFOV is 71.4 degrees. We outline a way to empirically test these results in Chapter 3. Based on the ratio of the blocked area to the total area of each lenslet, we determine that 75% of the light is blocked. We mounted all of the components using optical mounts on an optical table.

Using the main lens in reverse results in increased magnification and allows us to focus on objects placed very close (less than 1 cm) to the lens, which maximizes the angular range and perspective shift captured by the lenslets. However, designing a lighting system that stays clear of the camera yet appropriately illuminates the object is challenging. We used only direct overhead lighting on the object, which restricts us to using objects that can be effectively lit from above. We ensure the background was as dark as possible to reduce noise and spurious features in the reconstruction. To focus our image, we place the object as close to the main lens as possible, and then translate the sensor to the focal plane with respect to the object placement.

In a traditional imaging system with a single lens that uses the entire sensor, the resolution is maximized as each pixel on the sensor contributes directly to the final image. In our multiplexed light field camera system, multiple perspective images are projected onto the same sensor, leading to a decrease in resolution. However, the resolution decrease in our system is less significant compared to a traditional light field camera because our images overlap. The use of a large-aperture lens in our system introduces significant magnification, which impacts the depth of field. The aperture array helps mitigate some effects of the reduced DoF by limiting the aperture size of each sub-image formed by the lenslets (see Figure 2.2.

Our system can be modeled as an array of independent cameras that are optically overlapping. Each lenslet captures a slightly different perspective of the scene, contributing to a comprehensive light field capture. Importantly, the cameras do not share intrinsic parameters; they are functionally different cameras due to the warping introduced by the wide-angle lens. This configuration is not shift-invariant, making it difficult to calibrate using traditional methods for linear shift-invariant systems.

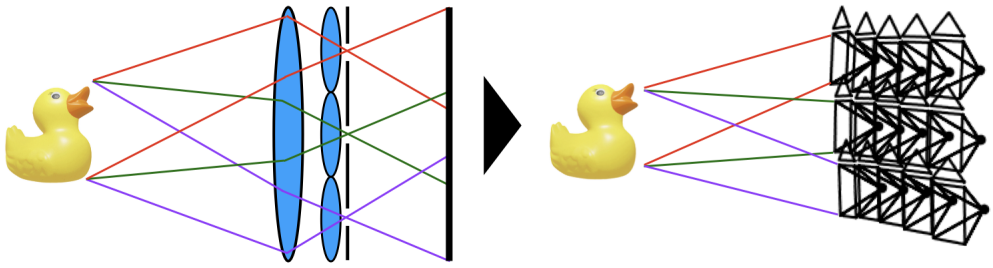


Figure 2.3: Each lenslet in the array functions as an individual camera, capturing a slightly different, overlapping perspective of the scene. This setup is analogous to an array of cameras that collectively capture a comprehensive light field.

Chapter 3

Novel View Synthesis for Multiplexing

3.1 Camera Calibration

Since we are modeling our system as an array of overlapping camera views, our reconstruction algorithm requires accurate estimation of the intrinsic and extrinsic parameters for each view. We employ a version of the standard structure-from-motion (SfM) pipeline using COLMAP [20, 21], initialized with DUSt3R [26], a deep learning-based method to predict 3D point clouds from single images.

Our calibration procedure involves capturing images through individual lenslets to isolate each view. We achieve this by covering all other lenslets and capturing images through each lenslet individually. This process is repeated for all lenslets in the array, effectively simulating the capture of the scene from multiple, slightly different viewpoints. This method ensures that each image corresponds to the perspective from a single lenslet. Additionally, we capture the effective field of view (FOV) of each lenslet by using a large light source in front of the system, which helps define the boundaries of each lenslet's coverage. Since we have 7 x 6 lenslets, this results in 42 calibration images and an additional 42 sub-view extent images.

The captured images are processed using COLMAP, a robust and accurate SfM pipeline commonly used in neural radiance field (NeRF) methods. COLMAP uses scale-invariant feature transform (SIFT) keypoints extracted from each image to match the same points between different views. SIFT keypoints are invariant to scale and rotation, making them suitable for matching across images with slight viewpoint changes. The matched keypoints provide the basis for reconstructing the 3D scene structure and refining the camera parameters through bundle adjustment. Since our scenes typically involve small objects and the individual lenslet images occupy a small portion of the input image, preprocessing is necessary. We crop the images to the region of interest before the usual downsampling step in order to enhance feature detection and matching on the non-sparse areas, and then apply the inverse operation on the reconstructed camera parameters. It is important that intrinsic parameters are not shared across each of the calibration images, unlike NeRF reconstruction from video, because each lenslet may have different intrinsic parameters due to manufactur-

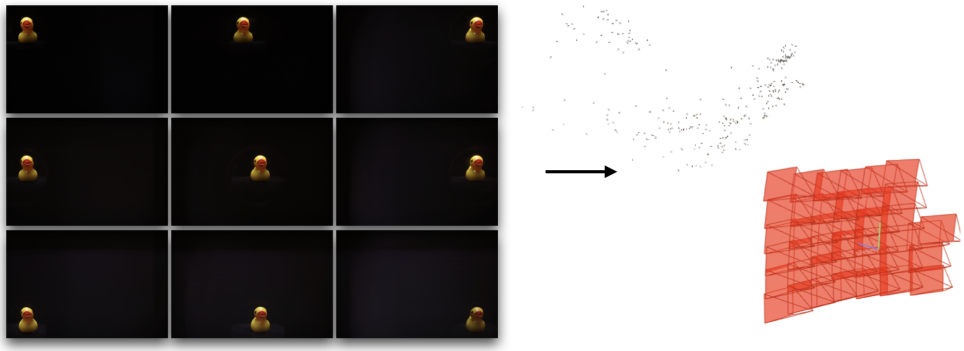


Figure 3.1: COLMAP reconstruction result showing the estimated camera poses and sparse 3D point cloud from calibration images. The calibration images shown are a subset of the 42 images used. The sparse point cloud indicates the rough 3D structure of the scene.

ing variations and optical distortions. Figure 3.1 shows a sample COLMAP reconstruction result from calibration images, indicating the estimated camera poses and sparse 3D point cloud, where the red frustums represent the reconstructed camera poses.

COLMAP relies on good initialization to converge to an accurate solution, especially in the absence of metadata such as focal lengths. There are failure modes associated with COLMAP, particularly with symmetric objects. The symmetry can create ambiguities in the location of camera views, leading to flattened or compressed point clouds, as illustrated in Figures 3.2 (a) and (b). COLMAP creates an initial pair of images and then iteratively refines the solution with additional images through bundle adjustment. Without good initialization—due to a lack of features in the image—COLMAP's performance is unreliable. To improve the initialization, we employ DUSt3R, a deep learning method that predicts dense 3D point clouds iteratively from pairwise images. DUSt3R leverages learned priors from large datasets to provide a rough estimate of the scene structure and camera poses. Although DUSt3R is less precise than COLMAP, it is usually able to provide a stable starting point for optimization. We extract the focal lengths and initial camera poses from the DUSt3R estimation and use them to initialize COLMAP's estimation. We find that this significantly improves the convergence rate and accuracy of COLMAP compared to random initialization.

Another way to measure the AFOV when we have the predicted 3D points and camera poses from COLMAP is to select a point in the 3D scene and measure the angle between the most extreme camera views that observe that point. Specifically, for a given 3D point, we compute the vectors from the point to the camera centers of the leftmost and rightmost cameras that see it. The angle between these two vectors represents the angular field of view for that point. Using this method on one of our COLMAP reconstructions in Chapter 4.2, we

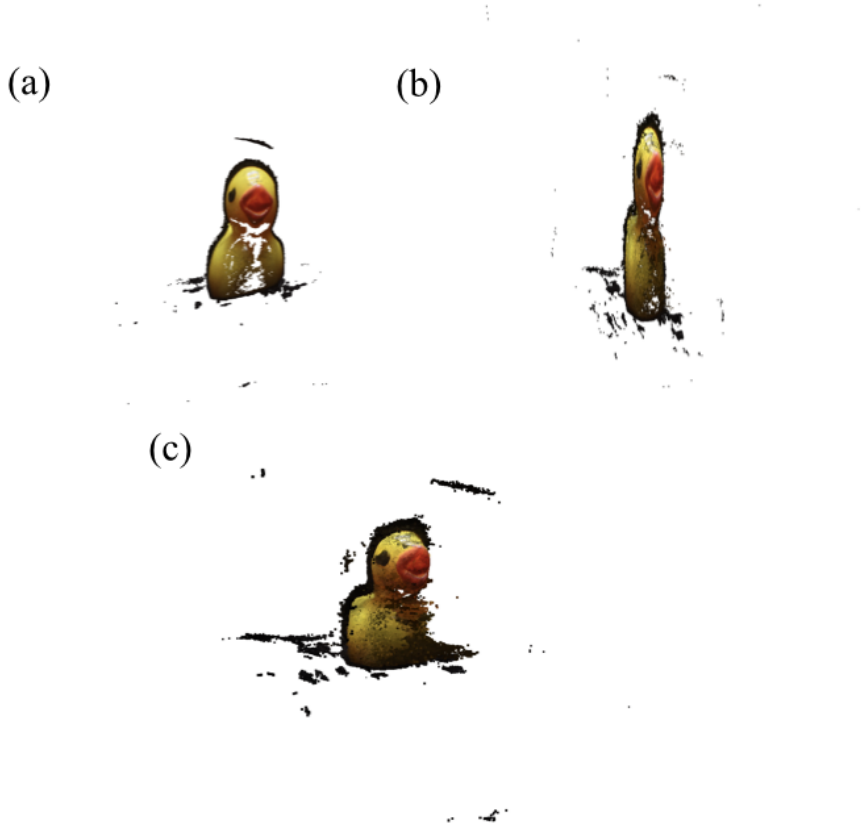


Figure 3.2: (a-b) COLMAP failure modes. Due to the symmetry of this object, there exists some ambiguity in the location of the camera views, leading to point clouds that are flattened or compressed. (c) shows a successful reconstruction.

measured an AFOV of 90.62 degrees between the most extreme cameras for the closest SfM point. This provides empirical validation of our system's ability to capture a wide angular range.

One significant advantage of our imaging system is that the calibration process only needs to occur once. Unlike traditional novel view synthesis techniques that require recalibration for each new scene, our system maintains its calibration across different objects. This reduces the computational load and improves the reliability of the reconstruction results. It also means that we can calibrate based on a known object and then reconstruct an object that would be difficult to estimate with COLMAP, such as objects with few features or challenging textures.

3.2 Gaussian Splatting Optimization

Our reconstruction pipeline builds upon Gaussian Splatting [7], the state-of-the-art technique for novel view synthesis. Gaussian Splatting models the scene as a continuous field of oriented 3D Gaussians, a flexible and expressive scene representation capable of representing our objects with a relatively small number of Gaussians—on the order of 20,000. The 3D Gaussians are initialized using the sparse point cloud produced from the SfM process in COLMAP. During optimization, the algorithm refines the Gaussians' 3D positions, opacities, anisotropic covariances, and spherical harmonic coefficients for color representation. In addition, adaptive density control steps add or prune Gaussians based on if they require densification or if their transparency is too low, respectively. A key contribution of this method is the fast and differentiable rendering process, which is designed to fully utilize GPU operations to enable fast training speed and real-time rendering. It involves splitting the screen into 16 x 16 tiles, projecting the Gaussians onto the 2D image plane, sorting, and applying standard α-blending.

Our solution adapts this technique to work with multiplexed light field images captured by our system, enabling reconstruction from a single image. The primary challenge in adapting Gaussian Splatting for our system is the need to handle multiplexed images. To address this, we modify the image formation model in Gaussian Splatting to simulate the multiplexed image formation by sampling the Gaussian representation at multiple viewpoints corresponding to the calibration data. Instead of rendering one image for each training pass, we render one image from each viewpoint in the camera and then combine them to create the multiplexed image. We create a composite image by integrating the sampled views, simulating the multiplexed image formation process of our hardware light field camera.

An inherent issue with our multiplexing approach is the ambiguity regarding which lenslet a particular ray originated from, especially in the overlapping regions. This can result in artifacts such as multiple copies of the object in the final Gaussian Splatting render. To mitigate this, we captured the full view extent of each lenslet by covering all other lenslets and using a large light source. This allowed us to define the valid regions for each lenslet's contribution. During training, we apply an out-of-bounds regularization to penalize Gaussians that contribute to regions outside the expected FOV of each lenslet. Specifically, after rendering each lenslet's image, we extract the pixels that fall outside the known circular FOV for that lenslet and apply an $\mathcal{L}_1$ loss to these pixels. This encourages the model to suppress contributions in out-of-bounds areas, reducing ambiguity. In addition, we add a 3D total variation (TV) regularizer to smooth the spatial distribution of Gaussians for higher image fidelity.

The overall loss function $\mathcal{L}$ used during our training process is thus defined as follows:

$$\mathcal{L} = (1 - \lambda_1)\mathcal{L}_1 + \lambda_1\mathcal{L}_{\text{D-SSIM}} + \lambda_2\mathcal{L}_{\text{FOV}} + \lambda_3\mathcal{L}_{\text{TV}}. \tag{3.1}$$

The 3D TV regularization is defined as follows:

$$\mathcal{L}_{\text{TV}} = \sum_{i,j,k} \sqrt{\left(v_{i+1,j,k} - v_{i,j,k}\right)^2 + \left(v_{i,j+1,k} - v_{i,j,k}\right)^2 + \left(v_{i,j,k+1} - v_{i,j,k}\right)^2} \tag{3.2}$$

where $v_{i,j,k}$ represents the Gaussian parameters at voxel position (i, j, k).

In the original Gaussian Splatting implementation, there is an operation that resets the opacity of the Gaussians after a few thousand iterations to prevent overfitting. However, we observed that this reset mechanism led to poor reconstruction performance in our setup. Thus, we disabled the opacity reset operation, allowing the model to better preserve the fine details in the multiplexed images.

Chapter 4

Experimental Results

Note: this chapter presents work done jointly with Vi Tran.

4.1 Simulation Experiments

To evaluate the effectiveness of our multiplexed light field camera system, we conducted a series of simulation experiments. These experiments aimed to validate our approach in a controlled environment where ground truth data is available. We used the Lego scene from the Blender dataset [14] to generate synthetic scenes with known camera parameters and viewpoints. The synthetic scenes were generated by positioning multiple virtual cameras around the Lego scene and capturing images from these viewpoints. We then synthetically combine these images into a single composite image as input for Gaussian Splatting. During the training step, we query images from every sub-view camera position and combine the images to calculate the loss with the training image. This approach allowed us to control the level of multiplexing by adjusting the number of virtual cameras and the degree of overlap between their images. We trained the algorithm on an RTX 3090 GPU.

In our simulations, we varied the level of multiplexing from 1.0 to 3.5 rays per pixel. We quantitatively evaluated the reconstruction quality using Peak Signal-to-Noise Ratio (PSNR), a standard metric used to measure the quality of reconstructed images compared to the ground truth, with higher PSNR values indicating better reconstruction fidelity. Our results demonstrated that even with significant multiplexing, our method outperformed Gaussian splatting with a single view. Figure 4.1 shows the performance comparison between single-lens and multilens cameras on the lego scene. Specifically, the advantage of the multilens system is most pronounced around 2.0 rays per pixel, which corresponds to a 100% increase in resolution. It is important to note that the rays per pixel metric may not be representative when based on the field of view of the system, as a more sparse object can tolerate more multiplexing when the actual image overlap is less. An additional multiplexing metric is taking the rays per pixel on the captured images, which is discussed in more detail in Chapter 4.2.

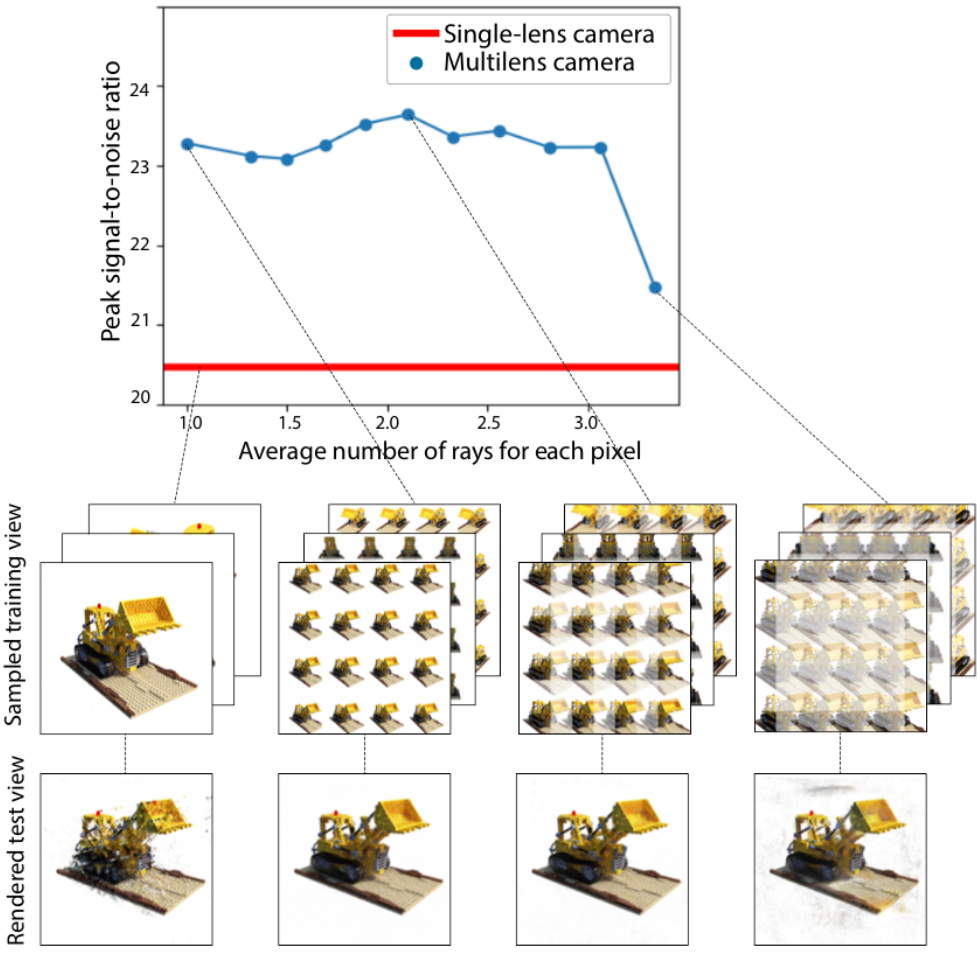


Figure 4.1: Performance comparison between single-lens and multilens cameras in simulation on the Lego scene. The multilens camera consistently outperforms the single-lens camera, achieving higher PSNR values, particularly around 2.0 rays per pixel.

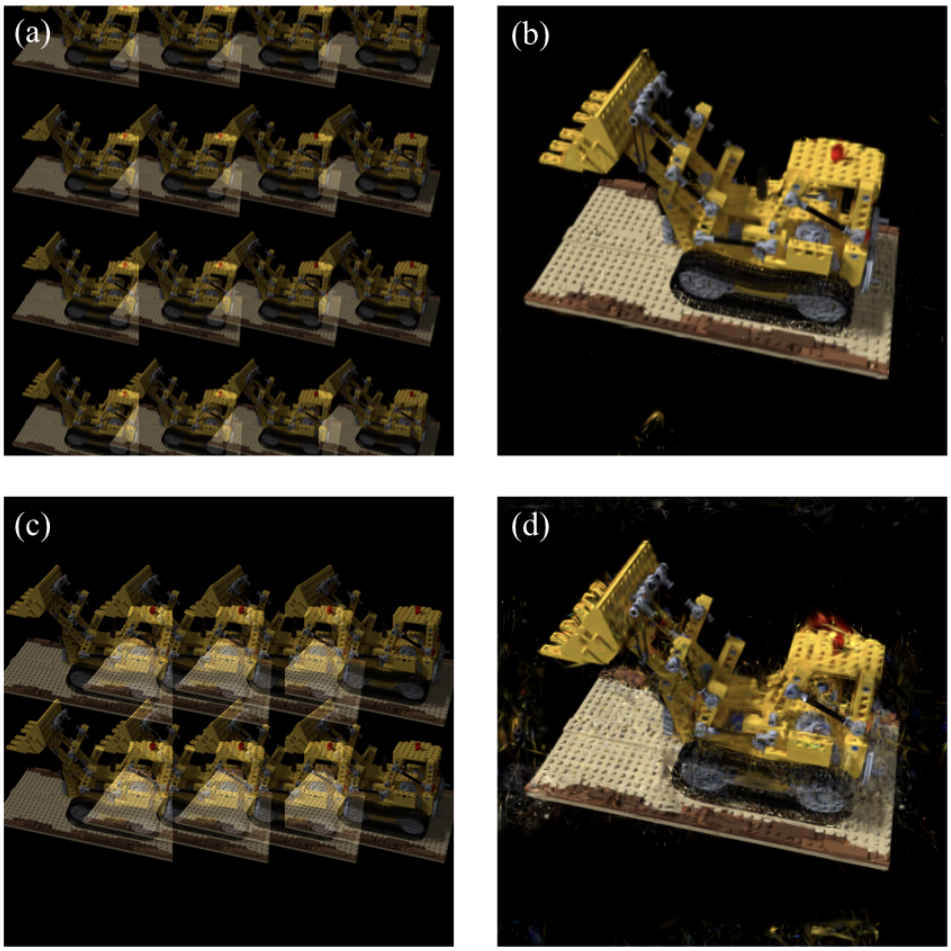


Figure 4.2: Synthetic reconstruction results with different amounts of multiplexing: (a) and (b) show the raw composite image and the reconstruction result at 1.5 rays per pixel, respectively. (c) and (d) show the raw composite image and the reconstruction result at 2.0 rays per pixel. The images demonstrate that higher levels of multiplexing lead to increased artifacts in the reconstructed scenes.

At higher levels of multiplexing, such as above 2.5 rays per pixel, artifacts begin to appear in the reconstructed images. These artifacts are primarily due to the increased ambiguity in the source of rays, making it challenging to accurately disentangle the multiple views effectively Figure 4.2 illustrates the synthetic results with different amounts of multiplexing. Thus, while our multiplexed light field camera system can handle moderate levels of multiplexing effectively, there exists a trade-off between the amount of multiplexing and the reconstruction quality—although a scene with higher sparsity may tolerate higher level of

multiplexing.

4.2 Real-World Experiments

Building upon the simulation results, we conducted a series of real-world experiments using the multiplexed light field camera system described in Chapter 2.2. We captured a multiplexed image and 42 calibration images of a small cluster of Nerds candy, as well as 42 images of each of the sub-view FOVs by placing a full-field light in front of the system. We then evaluated the reconstruction performance using our modified Gaussian Splatting pipeline.

We calculated the amount of multiplexing for our real-world system in two ways. First, we calculated amount of overlap using Equation 2.1 and using the sub-view FOV images for the lenslet contributions, leading to 6.5 rays/pixel. However, if we use the calibration images for the lenslet contributions, the multiplexing is actually 1.9 rays/pixel, indicating that the object actually takes up a very small amount of the full field of view.

In the first set of experiments, we used the calibration images as input, effectively simulating a scenario with no multiplexing. Figure 4.3 shows the results of these experiments. The PSNR for the no multiplexing scenario was measured at 30.04. The calibration renderings demonstrated high-quality results for a wide range of front-facing angles. However, beyond a 90 degrees rotation of the object, the reconstruction quality begins to degrade.

Next, we trained on only the multiplexed image. We obtained camera parameters for each of the sub-views by running the COLMAP SfM pipeline with the calibration images. Using the calibrated camera poses, we trained on one image using our modified Gaussian Splatting model. When training with the multiplexed, we observed a maximum PSNR of 26.35 with multiplexing. However, despite the relatively high PSNR, many artifacts and issues were apparent in the results. The multiplexed images exhibited problems such as smeared results (Figure 4.4a), double images (Figure 4.4b), and out-of-view renderings (Figure 4.4c). It is also worth noting that in all of the above cases, the volumetric reconstruction in the real-world experiments did not fully converge. These issues likely stem from mismatch between our model assumptions and the actual physical system. It is likely that with further tweaking of the reconstruction software, these results will improve to match the quality of the simulation results.

One idea for improving reconstruction quality is to add a total variation (TV) regularization term to views from non-captured sides to maintain structural stability. By applying TV regularization on far-away views, we can encourage the model to produce smoother and more coherent reconstructions, potentially mitigating artifacts caused by insufficient or ambiguous data. Another idea is to revisit the opacity reset that we disable in the training algorithm, or increase the density learning rate so that the algorithm can learn to prune unnecessary Gaussians at a higher rate.

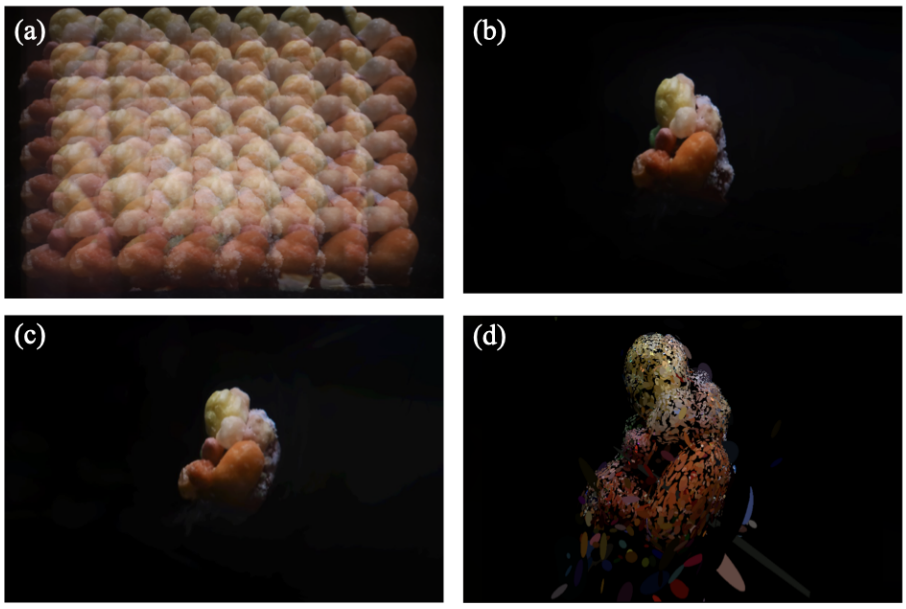


Figure 4.3: (a) Raw multiplexed image captured by our light field camera system. The image shows multiple overlapping views of the scene, each slightly shifted in perspective. (b-c) Gaussian Splatting reconstruction results at 0 degrees and 60 degrees from the optical axis, respectively. (d) Volumetric visualization of the Gaussians at full opacity and 10% size.

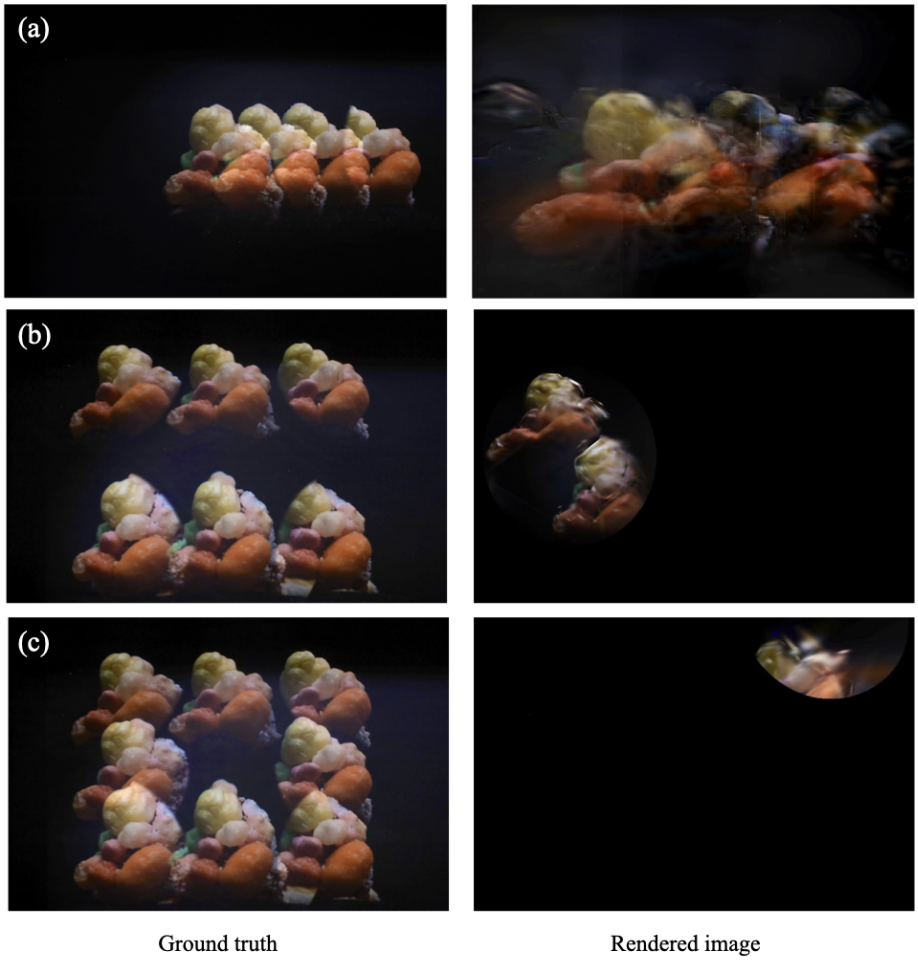


Figure 4.4: Test view renderings of the real-world reconstruction with multiplexing: (a) Results with high multiplexing, showing some smearing due to overlapping perspectives. (b) Double rendering with less multiplexing, indicating multiple object instances. (c) Out-of-view rendering, where parts of the scene appear outside the expected field of view.

Chapter 5

Conclusion

This work presents a novel approach to single-shot view synthesis using a multiplexed light field camera. Our camera system is designed to capture datasets for novel view synthesis much more quickly than traditional single-lens cameras, making it particularly practical for space-time methods to reconstruct videos from scenes. In simulation experiments, our system demonstrated superior performance compared to traditional single-view methods. Using the Blender dataset, we achieved higher PSNR scores than traditional single-lens systems, indicating that scenarios exist where multiple views achieve higher fidelity both with and without multiplexing. With multiplexing, we can physically capture more than 70 degrees of AFOV, and accurately reconstruct more than 120 degrees from one capture, showcasing the system's potential to capture higher resolution in a single shot. Our physical system achieves 1.9 rays/pixel of multiplexing, a 90% increase in pixel information over a light field imaging system with no overlapping, and we demonstrate higher-quality reconstructions on synthetic scenes with up to 2.5 rays/pixel of multiplexing when compared to both traditional light field images as well as monocular Gaussian Splatting.

In real-world experiments, we successfully captured multiplexed images of various small objects and scenes. The results indicate that we can achieve very good reconstructions without multiplexing. However, the multiplexed images currently suffer from artifacts, which we believe are due to an implementation bug rather than a fundamental limitation of our imaging system. One primary challenge is dealing with inherent ambiguities in overlapping images, especially when capturing symmetric objects. This ambiguity can lead to artifacts in the reconstructed images, as the system struggles to distinguish between different overlapping views.

The performance of our system in real-world scenarios is sometimes limited by the quality of the calibration process and the precision of pose estimation. While COLMAP provides robust pose estimation, it can be sensitive to the quality and features of the captured images, particularly for small objects. We addressed this limitation by using DUSt3R initialization, which improves pose estimation consistency.

Future work will focus on several key areas to further enhance the capabilities and applications of our multiplexed light field camera system. We will investigate the model mismatch

issues we are currently experiencing to identify and resolve the implementation bugs causing artifacts in the multiplexed images. By leveraging the fast capture capabilities of our system, we aim to extend our method for 3D video capture and reconstruction. This will involve addressing the challenges of temporal coherence and handling the additional complexity introduced by moving objects. We also plan to explore multi-shot approaches to achieve 360-degree 3D reconstructions with minimal captures.

Our system has promising applications in microscopy, macro photography, and 3D object reconstruction. Due to the large amount of magnification, our system is best suited for mesoscale imaging. It serves as a companion to view synthesis methods that predict other angles and feed into Gaussian Splatting, augmenting the abilities of single-view cameras to provide more effective information in one shot. The core contribution of this work is the development of hardware additions that improve the amount of scene information captured in a single shot when compared to monocular cameras, enabling much greater efficiency when using novel view synthesis techniques.

Bibliography

[1] E.H. Adelson and J.Y.A. Wang. "Single lens stereo with a plenoptic camera". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.2 (1992), pp. 99–106. DOI: 10.1109/34.121783.

[2] E.H. Adelson et al. *The Plenoptic Function and the Elements of Early Vision*. M.I.T. Media Lab Vision and Modeling Group technical report. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991. URL: https://books.google.com/books?id=w7FFHQAACAAJ.

[3] Emmanuel J. Candes and Michael B. Wakin. "An Introduction To Compressive Sampling". In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 21–30. DOI: 10.1109/MSP.2007.914731.

[4] Yang Fu, Ishan Misra, and Xiaolong Wang. "MonoNeRF: Learning Generalizable NeRFs from Monocular Videos without Camera Poses". In: *arXiv preprint arXiv:2210.07181* (2022).

[5] Todor Georgiev and Chintan Intwala. "Light field camera design for integral view photography". In: *Adobe System, Inc., Technical Report* (2006), p. 1.

[6] Changliang Guo et al. "Fourier light-field microscopy". In: *Opt. Express* 27.18 (2019), pp. 25573–25594. DOI: 10.1364/OE.27.025573. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-27-18-25573.

[7] Bernhard Kerbl et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". In: *ACM Transactions on Graphics* 42.4 (2023). URL: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[8] Marc Levoy and Pat Hanrahan. "Light field rendering". In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. New York, NY, USA: Association for Computing Machinery, 1996, pp. 31–42. ISBN: 0897917464. DOI: 10.1145/237170.237199. URL: https://doi.org/10.1145/237170.237199.

[9] Fanglin Linda Liu et al. "Fourier DiffuserScope: single-shot 3D Fourier light field microscopy with a diffuser". In: *Opt. Express* 28.20 (2020), pp. 28969–28986. DOI: 10.1364/OE.400876. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-28-20-28969.

[10] Andrew Lumsdaine, Todor Georgiev, et al. "Full resolution lightfield rendering". In: *Indiana University and Adobe Systems, Tech. Rep* 91 (2008), p. 92.

[11] Michael Lustig et al. "Compressed Sensing MRI". In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 72–82. DOI: 10.1109/MSP.2007.914728.

[12] Priyanka Mandikal and Venkatesh Babu Radhakrishnan. "Dense 3D Point Cloud Reconstruction Using a Deep Pyramid Network". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2019, pp. 1052–1060. DOI: 10.1109/WACV.2019.00117.

[13] Lars Mescheder et al. "Occupancy Networks: Learning 3D Reconstruction in Function Space". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[14] Ben Mildenhall et al. "NeRF: Representing scenes as neural radiance fields for view synthesis". In: *The European Conference on Computer Vision (ECCV)*. 2020.

[15] Ren Ng et al. "Light field photography with a hand-held plenoptic camera". PhD thesis. Stanford university, 2005.

[16] Jeong Joon Park et al. "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[17] Nicolas C. Pégard et al. "Compressive light-field microscopy for 3D neural activity recording". In: *Optica* 3.5 (2016), pp. 517–524. DOI: 10.1364/OPTICA.3.000517. URL: https://opg.optica.org/optica/abstract.cfm?URI=optica-3-5-517.

[18] Albert Pumarola et al. "D-NeRF: Neural Radiance Fields for Dynamic Scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[19] Sara Fridovich-Keil and Alex Yu et al. "Plenoxels: Radiance Fields without Neural Networks". In: *CVPR*. 2022.

[20] Johannes Lutz Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[21] Johannes Lutz Schönberger et al. "Pixelwise View Selection for Unstructured Multi-View Stereo". In: *European Conference on Computer Vision (ECCV)*. 2016.

[22] G. Scrofani et al. "FIMic: design for ultimate 3D-integral microscopy of in-vivo biological samples". In: *Biomed. Opt. Express* 9.1 (2018), pp. 335–346. DOI: 10.1364/BOE.9.000335. URL: https://opg.optica.org/boe/abstract.cfm?URI=boe-9-1-335.

[23] Vincent Sitzmann et al. "Implicit Neural Representations with Periodic Activation Functions". In: *Proc. NeurIPS*. 2020.

[24] Cheng Sun, Min Sun, and Hwann-Tzong Chen. "Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction". In: *CVPR*. 2022.

[25] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. "Splatter Image: Ultra-Fast Single-View 3D Reconstruction". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[26] Shuzhe Wang et al. *DUSt3R: Geometric 3D Vision Made Easy*. 2023. arXiv: 2312.14132 [cs.CV].

[27] Bennett Wilburn et al. "High performance imaging using large camera arrays". In: *ACM Trans. Graph.* 24.3 (July 2005), pp. 765–776. ISSN: 0730-0301. DOI: 10.1145/1073204.1073259. URL: https://doi.org/10.1145/1073204.1073259.

[28] Alex Yu et al. "pixelNeRF: Neural Radiance Fields from One or Few Images". In: *CVPR*. 2021.