Neural Field View Synthesis with Light Field Cameras

DAVID YOUNG

B.Eng (Hons)

Supervisor: Dr. Donald Dansereau

A thesis submitted in fulfilment of
the requirements for the degree of
Bachelor of Engineering (Honours)

School of Aerospace, Mechanical and Mechatronic Engineering
Faculty of Engineering
The University of Sydney
Australia

6 November 2022

Abstract

Novel view synthesis is an important research field which addresses the task of leveraging a set of source views of a scene to predict novel and unseen target views. Solving this has relevance towards a vast range of applications including cinematography, virtual reality, computer graphics, and robotics. However, existing view synthesis methods often degrade in performance across non-ideal and challenging scenario, such as when very few or limited source views are available.

To improve performance in these scenarios, we propose using light field cameras, which capture rich angular information of a scene, to replace conventional camera imaging. We develop a view synthesis pipeline for light field images by leveraging neural radiance fields (NeRF), a technique which learns a scene representation using a neural network, and demonstrate the reconstruction performance advantages of our approach over conventional imaging pipelines. Furthermore, we apply light field subsampling techniques to optimise our pipeline for both reconstruction quality and computational cost. Ultimately, we find that leveraging the information contained in light field images offers performance benefits over conventional imaging.

Overall, the contributions of this work can be utilised towards applications such as enabling higher quality and lower cost camera setups for complex cinematography and virtual reality techniques, as well as facilitating higher performance in robotic navigation by enabling more robust and accurate understandings of scene geometry.

Acknowledgements

I would like to thank my supervisor, Dr. Donald Dansereau, for his guidance and insight.

Statement of Contribution

- I conducted the background research and literature review of relevant work.
- I wrote code to perform preprocessing on an existing light field image dataset collected by Digumarti *et al.* [**1**] to repurpose the dataset for our experiments.
- I designed and implemented in code the light field ray sampling approaches proposed in this work.
- I used the existing implementation of NeRF from NVIDIA's InstantNGP [**2**] and incorporated it in our light field image view synthesis pipeline.
- I designed and carried out the experiments to evaluate our pipeline and ray sampling approaches.
- The discussions and conclusions in this work are my own, with insight from discussions with my supervisor.

David Young

Acronyms

MLP: multilayer perceptron.

NeRF: neural radiance field.

PSNR: peak signal-to-noise ratio.

SfM: structure from motion.

SSIM: structural similarity index measure.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Understanding a scene's appearance and structure from few or limited observations is a challenging yet prevalent computer vision problem. Novel view synthesis is a rapidly growing research field which addresses this challenge: given a number of source views of a scene, a view synthesis task aims to reconstruct a novel view of the scene from a previously unseen perspective. Solving this problem offers numerous applications in areas such as computer graphics, cinematography, virtual reality and robotics.

For example, in robotics applications, view synthesis can be applied to generate or predict unseen views which may facilitate more effective operation within a challenging environment. In cinematography, a classic application of novel view synthesis is the 'bullet time' effect from *'The Matrix (1999)'*, which depicted a slow-motion revolving camera shot of a scene appearing almost frozen in time. This effect was achieved using a filming setup of 120 cameras recording the scene simultaneously from varying perspectives. In post-production, view synthesis techniques were employed to interpolate novel perspectives of the scene unseen by the recorded footage, ultimately producing a seamless orbiting film shot of the scene. With the significant advances in view synthesis techniques across recent years, the same effect would likely be achievable in the present day with fewer cameras views of the scene. Generally, it is desirable to perform view synthesis with fewer views, as this enables lower-cost imaging setups and operation using less measured data of the scene.

However, current state-of-the-art view synthesis methods still exhibit inherent flaws and performance drawbacks in more challenging contexts where very few camera views are available. In this work, we aim to address this, and enable higher quality view synthesis reconstructions with less measured data of the scene.

1.2 Knowledge Gaps

Novel view synthesis embodies a vast field of existing research and techniques. Many state-of-the-art view synthesis methods which achieve the highest performance leverage neural scene representations, which represent and encode the details of a scene's appearance and geometry into the weights of a neural network, which can then be used to generate novel views once the neural network is trained. In particular, many methods using neural radiance fields (NeRFs) [**3**], one type of neural scene representation, have achieved state-of-the-art performance in novel view synthesis benchmarks on public datasets such as ShapeNet [**4**] and RTMV [**5**].

However, NeRF and other state-of-the-art techniques often underperform in more challenging and non-ideal scenarios. One major challenge faced by these techniques is few-shot view synthesis: operating in contexts where very few or limited camera views are available. In these scenarios, view synthesis techniques are more prone to failure due to having insufficient information to synthesise novel views. To improve the reconstruction performance in these scenarios, we propose a view synthesis pipeline using light field images as input, which incorporates a neural scene representation to reconstruct novel views. Light field cameras are a type of imaging device which capture four-dimensional light field images which contain rich angular information of how the view of a scene changes with slight shifts in perspective.

Light field imaging has been leveraged in prior work to achieve performance gains over conventional camera imaging in tasks such as simultaneous localisation and mapping [**6**], visual odometry [**1, 7**] , visual servoing [**8**], and feature detection [**9**]. These tasks, which traditionally utilise conventional camera imaging, all demonstrate various performance advantages when light field imaging is used instead.

However, adapting an existing technique or pipeline which relies on conventional imaging to instead operate on light field images is often non-trivial. In the context of NeRF and other neural scene representations, an established approach to using light field image inputs does not currently exist. As NeRF trains on light rays measured in a camera image, one straightforward approach to adapting NeRF to train on light field images would be to simply input the full set of light rays contained in the light field image into the model. While this would likely achieve high reconstruction performance, the associated computational cost would also increase significantly due to the increased input size into the NeRF model. In this work, we investigate approaches to light field input which maintain high reconstruction performance without significant expense to computational cost. While light field images contain rich angular information of the imaged scene, much of the information contained in a light field image is duplicated and redundant. Hence, we exploit this property of light field images and investigate selecting a reduced subset of the overall measured ray data in a light field image to optimise for computational efficiency.

1.3 Contributions

The contributions of this thesis are:

- We develop a light field image view synthesis pipeline leveraging neural radiance fields (NeRFs), and demonstrate our pipeline to achieve better reconstruction quality and faster training convergence in comparison to a conventional image state-of-the-art NeRF pipeline.
- We develop various methods to subsample the light rays contained in a light field image to discard less relevant information, and demonstrate that we can use ray sampling to retain high reconstruction quality while training on less measured data of the scene.
- We perform a comparison on different ray sampling techniques by evaluating data fidelity, and show that certain light rays within a light field image are more useful than others for view synthesis tasks.

1.4 Thesis Outline

In Chapter 2, we present an overview of the background knowledge required to understand the contributions, methodologies and results of this work. In particular, we detail the fundamentals of light fields and light field cameras, and explain the key concepts and intuitions behind NeRF and how they can be applied for view synthesis.

Chapter 3 provides a review on relevant research literature, divided into three primary sections. First, we discuss current state-of-the-art methods for novel view synthesis which operate on conventional camera images. Following this, we detail view synthesis methods which are capable of operating on light field images. Finally, we explore existing methods for subsampling light field images, which aim to create sparse light fields which contain less redundant information. In this chapter, we highlight shortcomings and limitations in the existing research literature, and the resultant knowledge gaps which our work addresses.

In Chapter 4, we detail the key methodologies and ideas implemented by this work. We first overview our view synthesis pipeline, then present a more detailed explanation of each component in the pipeline. The various methods used for performing ray sampling are detailed.

Chapter 5 describes the experimental setup and implementation details used to realise the ideas described in the preceding chapter, and then analyses the results obtained from the experiments.

In Chapter 6, we present discussions of our experimental results, derive key insights and implications from our work, and discuss how our work addresses the knowledge gaps and limitations of the research literature.

Chapter 7 concludes our work. Here, we summarise the key contributions made, and the significance of this research. Finally, we highlight directions for future research which follow and build upon the contributions presented in this work.

CHAPTER 2

Background

2.1 Light Fields

Light travels through space in straight lines, or rays, where each ray is defined by a direction and radiance. Hence, all the light flowing through a scene can be modelled by knowing the radiance and direction of every light ray in the scene. The *light field* is a function which models this, mapping every possible light ray in space to the radiance observed by the ray. Hence, if the full light field of a scene is known, then the radiance of every light ray in the scene is known. The appearance of the scene from any view could then be generated by using the radiance values of the light rays which are observed by a given view. This process of generating views or images from a light field is referred to as light field rendering.

A frequently used parameterisation of the light field is the two-plane parameterisation, originally defined by Levoy and Hanrahan [**10**]. This definition parameterises the light field as a 4D function, which models each light ray direction by its intersection with two defined 2D planes in space, as shown in Figure 2.1. Referring to the two planes as the s-t plane and the

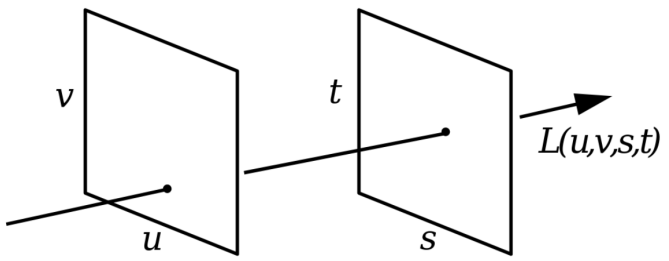

FIGURE 2.1. **The two-plane parameterisation of a 4D light field.** [**10**]

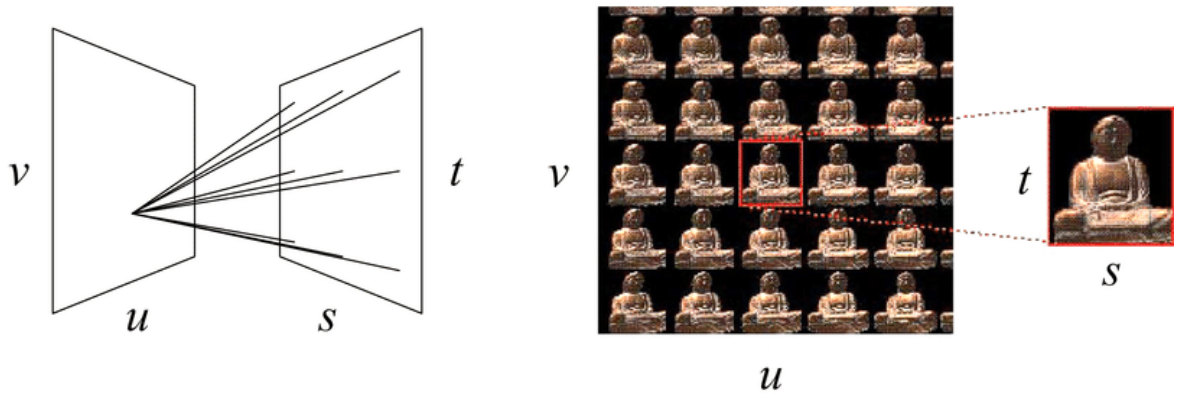

FIGURE 2.2. **Visualisation of a light field [10].** Consider a camera placed at a single point in the u-v plane, facing towards the s-t plane. The incoming light rays received by that camera view (indicated by the black lines between planes) capture a single image, which is outlined in red on the right. By choosing another point on the u-v plane to place a camera, an image with a slight perspective shift is attained. A light field camera can be imagined as a number of cameras placed at different points on the u-v plane, which capture a 2D grid of images similar to the image grid shown in this figure.

u-v plane, this characterises a four dimensional light field $L(u, v, s, t)$. The direction of each light ray is thus defined by a pair of 2D coordinates (u, v) and (s, t). The light field function L maps each light ray defined by the four parameters to the radiance observed by that ray.

2.2 Light Field Cameras

A light field camera, also known as a plenoptic camera, is an imaging device which images the light field of a scene. More simply, a light field camera can be imagined as a number of conventional cameras rigidly arranged in a grid or 2D array. Consequently, a light field image comprises an array of subviews, each comprising a conventional camera image. Figure 2.2 illustrates how a light field camera samples the light field of a scene, relating to the two-plane parameterisation of a 4D light field presented in Figure 2.1.

Consequently, a straightforward method of constructing a light field camera is simply using a 2D array of conventional cameras. This method of light field capture was originally proposed by Wilburn *et al.* [**11**]. Figure 2.3 depicts a light field camera built using this

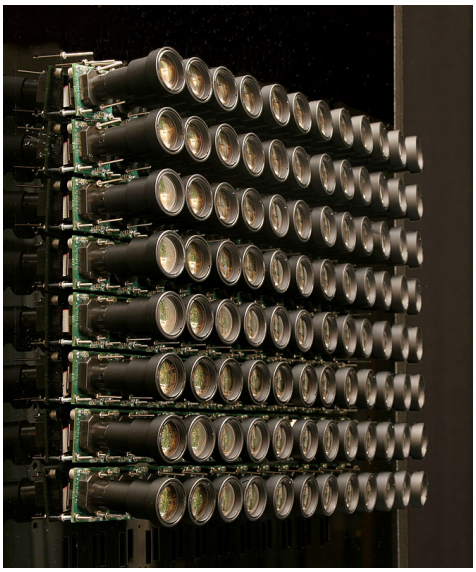

FIGURE 2.3. **A light field camera built from an array of monocular cameras [11].**

method. Alternatively, light field cameras can also be constructed using a lenslet array, a type of camera lens which allows for light field imaging with only a single camera sensor, as originally proposed by Ng *et al.* [**12**]. Lenslet arrays offer a lower cost solution to light field imaging as a result of the fewer camera sensors required. Additionally, conventional cameras can be repurposed to take light field images through the addition of a lenslet array, enabling higher accessibility. However, light field cameras built using camera arrays typically offer higher spatial resolutions, though they are more costly to build and require more processing requirements due to the many cameras used.

2.3 Neural Radiance Fields (NeRF)

2.3.1 Scene Representation

A neural radiance field (NeRF) [**3**] is a type of neural scene representation which learns the density and view-dependent radiance of every spatial location within a scene. To understand this definition, consider a single 3D point in space defined by $\mathbf{x} = (x, y, z)$. This point in the scene is observed to have some radiance, which determines the RGB colour of the point as

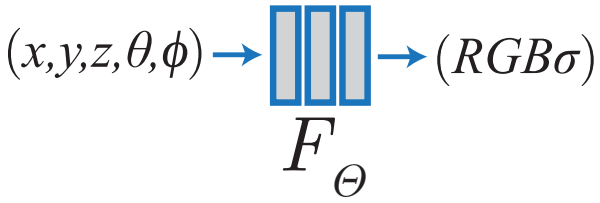

FIGURE 2.4. **The inputs and outputs of the NeRF model [3].** F_Θ is the multilayer perceptron (MLP) used to learn the scene representation.

seen by the observer. However, the observed radiance, or emitted colour, of a point in a scene is often dependent on the viewing direction. This is particularly evident for reflective surfaces which exhibit high specularity: a point on a mirror, for example, will typically look different, and thus have a varying radiance, depending on the direction it is observed from. Thus, a 2D viewing direction $\mathbf{d} = (\theta, \phi)$ is defined to parameterise the direction from which a point is observed. These two angles θ and ϕ correspond to the two angles associated with a direction in a 3D spherical coordinate system.

In addition to the view-dependent radiance of a point, each point in the scene also has a volume density denoted by σ. The density value assigned to a point represents the probability of a light ray terminating at that point in space. A point on a non-transparent solid surface or object would be associated with a higher density, while a point in free space would be associated with a lower density. Similarly, a point occupying some translucent medium or occlusion (e.g. smoke, fog) would have a density value proportional to the amount of light which can pass through the point. Unlike radiance, the density of a point is not dependent on viewing direction. This property guarantees multiview consistency, meaning that the scene geometry learnt by the model is consistent across different views of the scene.

The NeRF model defines a function F_Θ mapping from a point $\mathbf{x}$ and viewing direction $\mathbf{d}$ to the density σ of the point and the observed radiance $\mathbf{c}$ of the point when viewed from the viewing direction $\mathbf{d}$. The observed radiance is defined in terms of the emitted colour on the RGB spectrum, parameterised by $\mathbf{c} = (r, g, b)$. Figure 2.4 illustrates this 5D function. This can be more concisely expressed as:

$$F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$$

This 5D function mapping each $(\mathbf{x}, \mathbf{d})$ to $(\mathbf{c}, \sigma)$ is a scene representation known as a radiance field. Consequently, a *neural* radiance field leverages neural networks to learn and represent the radiance field of a scene.

2.3.2 Rendering Views

Rendering a view of a scene from a target camera pose involves determining the colour associated with all the light rays which are observed by the camera, then collating all the colours into an image, where each pixel corresponds to one observed light ray and its associated colour.

Radiance fields detail the density and radiance properties of specific spatial locations. A light ray passes through an infinite number of points in space arranged along a straight line. Therefore, the colour observed by a light ray is consequently a function of the radiance and density properties of all of the infinite points through which the ray passes. Using a radiance field to render the colour of a light ray passing through a scene requires classical volume rendering techniques for ray tracing through a volume density [**13**]. To render the colour seen by a light ray, the radiance and density properties of many discretely sampled points along the light ray are queried from the neural network. An integral of the radiance and density properties along the light ray is computed, which renders the single colour observed by that light ray. Therefore, rendering a single light ray requires many queries to the neural network scene representation.

2.3.3 Learning a Scene

NeRF uses a multilayer perceptron (MLP) neural network (denoted by F_Θ) to learn the 5D radiance field of a scene. A NeRF model trains on a set of light rays and their known colours. Therefore, the pixel information contained in the set of source camera images is first transformed into a set of rays, where each ray is labelled with its observed colour. The training process then selects a random ray and renders a predicted colour for that ray by applying volume rendering techniques as outlined above. A squared error loss is then computed

between the predicted pixel colour rendered by the model and the true pixel colour from the ground truth data. As the rendering process and loss function are entirely differentiable, the neural network can be subsequently trained using the classical machine learning optimisation techniques of backpropagation and gradient descent. Repeating this training process iteratively across a large number of rays allows the model to converge to an accurate representation of a scene's radiance field. This can then be used to synthesise novel views which were unseen among the set of source camera views.

CHAPTER 3

Literature Review

In this chapter, we provide a review on research literature relevant to our work. We highlight shortcomings and limitations in the existing research literature, and the resultant knowledge gaps addressed by our work.

3.1 View Synthesis from Conventional Imaging

Many methods exist to perform novel view synthesis on conventional images, leveraging various techniques ranging from simple interpolation to complex deep learning approaches.

One straightforward and naive approach to synthesise novel views is using light field interpolation techniques [**10, 14**]. These approaches construct a light field using the set of images and their known camera poses, and then use interpolation within the light field, resampling information from the set of source views to determine and reconstruct a novel target view. However, to achieve accurate view synthesis performance, these naive interpolation-based approaches require a dense cluster of source views in close vicinity to the desired target view. When there are insufficient source views nearing the target view to interpolate from, the performance drastically degrades.

A more recent view synthesis approach involves initialising a mesh representation of a scene and optimising this mesh to fit the source image views through gradient descent. Both Liu *et al.* [**15**] and Chen *et al.* [**16**] utilise this mesh optimisation approach in combination with a differentiable rendering process. From the mesh representation of a scene, predictions of the source views can be rendered. As the rendering process is differentiable, the loss between

the predicted image and its ground truth can be used for a gradient descent optimisation. However, these approaches which use mesh optimisation sometimes underperform when gradient descent converges to a non-optimal solution, or when the mesh representation is poorly initialised.

As an alternative to mesh representations, an alternative view synthesis approach uses volumetric representations of a scene. [**17, 18**]. Both Flynn *et al.* and Mildenhall *et al.* use a multiplane image representation (MPI), a volumetric scene represenation which approximates the scene as a stack of layered planes at various depths. By blending sampled views of from the MPI, a view from a previously unseen perspective can be reconstructed. One key limitation of volumetric representations is the discrete nature of the representation. As the representation is not continuous, these approaches scale poorly when higher resolution is desired.

3.1.1 Neural Radiance Fields NeRF

As detailed in the background chapter, NeRF is a state-of-the-art view synthesis technique which learns a scene representation using a neural network. This representation is continuous, and can be sampled at any arbitrary resolution, which addressed the limitations of many previous view synthesis methods. Across many benchmarks and datasets [**4, 5**], NeRF achieved state-of-the-art performance in view synthesis. However, the original NeRF implementation faced two key challenges. Firstly, the training time was extremely slow, taking from hours up to days to converge for a scene when training on a single GPU. Secondly, for the NeRF model to converge to an accurate scene representation, typically a large number of views sparsely distributed across the scene are required. When very few or limited images are provided, reconstruction quality suffers due to a lacking sufficient information of the scene. Many variations of NeRF have been introduced to address these challenges and offer improvements over the original NeRF implementation.

NVIDIA's InstantNGP [**2**] provides an implementation of NeRF which introduces multiresolution hash encoding. This technique utilises a hash table which maps spatial locations in the

scene to trainable feature vectors which are optimised using stochastic gradient descent. This approach speeds up the training process by several orders of magnitude, allowing for training to converge for a scene within the range of seconds to minutes.

Light Field Networks (LFNs) [**19**], an entirely different neural scene representation, learn the scene as a 4D light field instead of a 5D radiance field. By representing the scene as a light field, rendering a ray is achieved with a single query to the neural network, instead of the large number of queries required to render a ray in a radiance field using volume rendering methods. This allows for extremely fast training and rendering times, with the capability of running in real-time. However, despite the drastic speed advantage, a primary disadvantage of LFNs is that they are not multi-view consistent. In radiance fields, the view-independent density term guarantees consistent scene geometry across different views. In a 4D light field, this property does not exist. Instead, to learn a scene which is multi-view consistent, LFNs implement a supervised meta-learning training approach to learn a prior over light fields of other scenes.

Pixel-NeRF [**20**], another NeRF variation, introduces a convolutional encoder which trains across multiple scenes to develop a scene prior. By using supervised learning to develop a learned prior over many scenes, approaches such as LFNs and Pixel-NeRF are able to achieve much higher performance on view synthesis with very few views, addressing one of the key disadvantages of the original NeRF implementation. While supervised models can often achieve higher performance through leveraging prior knowledge of other scenes, they require large datasets of high quality data for training. Unsupervised view synthesis models, on the other hand, only leverage the information contained within a single scene, without exploiting prior knowledge of other scenes. Hence, unsupervised models do not require prior training on large datasets to form a prior. Ultimately, both supervised and unsupervised approaches are useful in differing contexts when these trade-offs are considered.

Currently, unsupervised view synthesis techniques exhibit poor performance in few-shot scenarios, defining the primary knowledge gap which our work aims to address. This poor performance is often a result of having insufficient information of the scene. While existing solutions in the literature manage to compensate for this information deficiency by

implementing supervised learning techniques across scenes to develop a model with prior knowledge, in this work we instead address the fundamental limits of conventional camera data capture, and investigate how light field imaging can offer advantages over conventional cameras for view synthesis tasks. By leveraging light field images, which image a scene in both angular and spatial detail, we aim to develop a view synthesis pipeline which is able to better exploit measured data of a scene.

3.2 View Synthesis from Light Field Images

Existing research literature on view synthesis from light field images primarily addresses synthesis of views within a single light field image, rather than synthesising new views from a set of light field images. Kalantari *et al.* [**21**] developed a supervised deep learning approach which uses only the four corner subviews of a light field image to accurately reconstruct any interpolated view within the light field image.

Since a light field image comprises a dense cluster of conventional camera images, view synthesis approaches which operate on conventional images can typically also be applied to light field images by naively using each of the views within the light field image as input. Hence, many of the view synthesis techniques outlined earlier could be used on light field images.

Overall, existing view synthesis methods which are capable of operating on light field images do not optimally exploit properties of light field camera imaging. In our work, we aim to more optimally leverage light field images to perform novel view synthesis from sparsely distributed source camera views for both multi-view and few-shot scenarios.

3.3 Sparse Light Fields

Though light field images provide significantly more angular information of the scene in comparison to a conventional camera image, much of the information contained in a light field image is redundant. This is due to each of the 2D image subviews within a light field

image capturing the same scene with only a slight shift in perspective between subviews. The presence of this redundant information allows light field images to be highly compressible: in 1996, Levoy and Hanrahan [**10**] demonstrated a compression system to compress light fields by over a factor of 100 with very minor fidelity loss. Since then, other compression approaches have achieved even higher compression rates on light field data.

An alternative approach to reducing redundant information in a light field image is to selectively choose some subset of the data contained within the light field image which contains the most useful information, and discard the rest of the data which contains less useful information. As light field images are comprised of a certain number of conventional camera views, a frequently used approach to constructing sparse light fields is to consider a reduced set of the subviews within the light field image.

Exploiting sparse samples of a light field has been adopted in many computer vision tasks. For the task of depth estimation from light fields, Jiang *et al.* [**22**] devised an approach to perform depth estimation using a small subset of the subviews in a light field image, and demonstrated comparable performance to other state-of-the-art algorithms which used the full set of light field views. As detailed earlier, for the task of view synthesis within a light field image, Kalantari *et al.* [**21**] developed a supervised deep learning approach which used only the four corner subviews from the light field image to accurately predict interpolations of the remaining views, indicating that a small sampling of a light field image's views can retain a large majority of the information in the image.

Instead of sampling subviews post-capture, a different approach involves implementing view sampling at the hardware level. For a light field camera constructed from a camera array, this simply involves removing certain cameras. The EPIModule is a sparse light field camera which implements this concept, only retaining cameras arranged in a cross pattern as shown in Figure 3.1. A light field image captured from the EPIModule still contains considerable angular information of the scene due to the various camera views, but with a reduced amount of redundant information due to the sparser sampling of views. Digumarti *et al.* [**1**] used the EPIModule camera to perform depth estimation and visual odometry with unsupervised deep

FIGURE 3.1. The EPIModule sparse light field camera.

learning techniques, demonstrating that as with traditional light field cameras, sparse light field cameras also retain advantages over conventional camera imaging.

Overall, existing approaches to constructing sparse light fields sample specific subviews from the full set of subviews within the light field. We aim to extend on this by evaluating the importance of subview positioning, and determining which subviews in a light field camera are most valuable and contain the most useful information for view synthesis tasks. Furthermore, we leverage ray sampling methods to analyse data fidelity of subsampled light fields on a per-ray basis, which has not been performed in existing literature. Leveraging light field subsampling on a per-ray basis instead of the per-view basis which existing literature techniques employ allows for a more fine-grained control for distinguishing between useful and redundant information contained within the light field. Through this, we aim to attain further insight into the most optimal methods to measure the light rays of a scene which maximise the amount of information retained, ensuring high data fidelity.

CHAPTER 4

Methodology

In this chapter, we describe the theory and key concepts of our light field image view synthesis pipeline. We first present a full overview of our pipeline, and then introduce various methods of performing ray sampling to optimise our pipeline.

4.1 Pipeline Overview

As with conventional view synthesis tasks, our pipeline requires both images and the associated camera poses for each view as input. However, our pipeline differs from a conventional view synthesis pipeline as it operates on a set of light field images instead of conventional camera images.

An overview of our pipeline is depicted in Figure 4.1. For each light field view, we perform subsampling to select a subset of all the light rays measured in the image, ultimately reducing the total measured information. There are two key reasons for performing this subsampling step. Firstly, as established in the background and literature review, light field images contain redundant information of the scene. Hence, it is likely that a smaller subset of the full image can be used to train a view synthesis model with minimal sacrifice in reconstruction quality. By training on a smaller input, we can reduce the computational cost in both time and memory. This relates to the second reason for ray sampling: to enable adjustment and optimisation of the trade-off between reconstruction quality and computational cost. This trade-off can be adjusted by varying the rate of sampling. If a high sampling rate is utilised, then a majority of the full light field data is used for training, which will produce higher reconstruction quality at the expense of computational cost. For a low sampling rate, reconstruction quality would

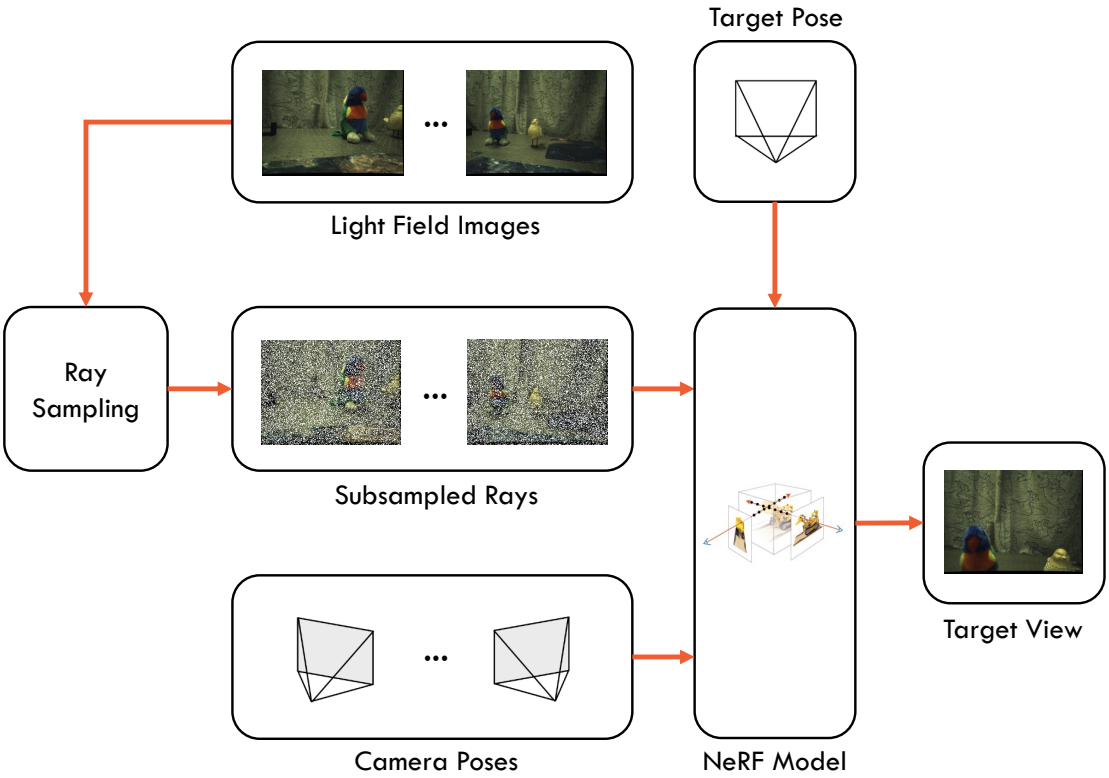

FIGURE 4.1. **Pipeline overview.** Our pipeline requires a set of light field images and the associated camera poses. The rays contained within each image are sampled using a ray sampling method, and then used to train a NeRF model to render novel views for a specified target pose.

be decreased, while training would be less computationally costly. However, it is unclear and non-trivial as to what the optimal method for ray sampling would be. We develop various methods to perform ray sampling, which are detailed later in this chapter.

The next step of our pipeline trains a NeRF model to learn a scene representation using the sampled rays from the ray sampling step and the camera poses for each of the light field images. Once training is complete, we can query target camera poses to the NeRF model to render novel views of the scene.

4.2 Ray Sampling

For any sampling method, we define a sampling rate s, a decimal value between 0 and 1 which expresses the proportion of light rays which should be selected for training. For example, if a light field image captures $100{,}000$ light rays and a sampling rate of $s = 0.5$ is used, the resulting sampled set should contain $50{,}000$ rays.

We explore many different ray sampling methods, with the goal of maximising data fidelity performance. An optimal sampling method will select the set of rays which result in the highest reconstruction quality for every sampling rate. We develop and evaluate four different sampling approaches: uniform, random, view-based, and image gradient sampling methods.

To better characterise and categorise the behaviour of different sampling methods, we define certain attributes which describe fundamental principles for sampling approaches. Firstly, each sampling method is classified as either fixed pattern or variable pattern. This attribute describes whether the sampling pattern varies across different source views of the scene. In fixed pattern approaches, a constant sampling pattern is employed across different views, such that if a pixel in one image is sampled, then the pixel at the same position in the image is sampled for every other image of the scene. In variable pattern approaches, the sampling pattern may vary across different views. This attribute has implications on computational efficiency, which will be discussed later.

We also define a *semantic sampling* attribute, which classifies whether a sampling method leverages the image information to inform the sampling. In other words, for a sampling method which samples semantically, the resultant sampling of an image is a function of the pixel values contained within the image. A semantic sampling method must also be a variable pattern approach, as a method which samples based on the image context must by implication be able to sample using varying context-dependent patterns across different images.

In addition to the four sampling methods defined above, we also evaluate fixed pattern variations for two of the methods which use variable pattern sampling. We introduce *fixed random sampling* and *fixed view-based sampling*, which are the respective fixed pattern

variations of random sampling and view-based sampling. An overview of the sampling methods implemented in this work are shown in Table 4.1.

TABLE 4.1. **Sampling method attributes.**

Sampling Method	Fixed/Variable Pattern	Semantic Sampling
Uniform	Fixed	
Random	Variable	
View-Based	Variable	
Image Gradient	Variable	✓
Fixed Random	Fixed	
Fixed View-Based	Fixed	

4.2.1 Uniform Sampling

We first establish the uniform sampling approach, which samples uniformly in a fixed pattern across each subview in the light field image, such that the selected rays have a evenly structured spatial distribution, where every sampled ray is equidistant from its sampled neighbours. This sampling method is similar to naively downsampling the spatial resolution of each subview. Thus, uniformly sampling ensures the sampled rays are uniformly distributed across the image space. This sampling method should perform well if all spatial regions of the captured image are equally important towards maximising reconstruction quality. An example of a sampled set of rays using the uniform sampling method is shown in Figure 4.2.

4.2.2 Random Sampling

Random sampling samples rays randomly from a uniform statistical distribution across each light field subview. Hence, the behaviour resembles the uniform sampling method, but with an added element of randomness which guarantees variance in the pixel positions which are being sampled. In uniform sampling, the pixels in the same positions are sampled across every light field view.

There are two methods in which random sampling can be performed on a light field image. For a sampling rate s and a light field image containing N rays, we aim to construct a

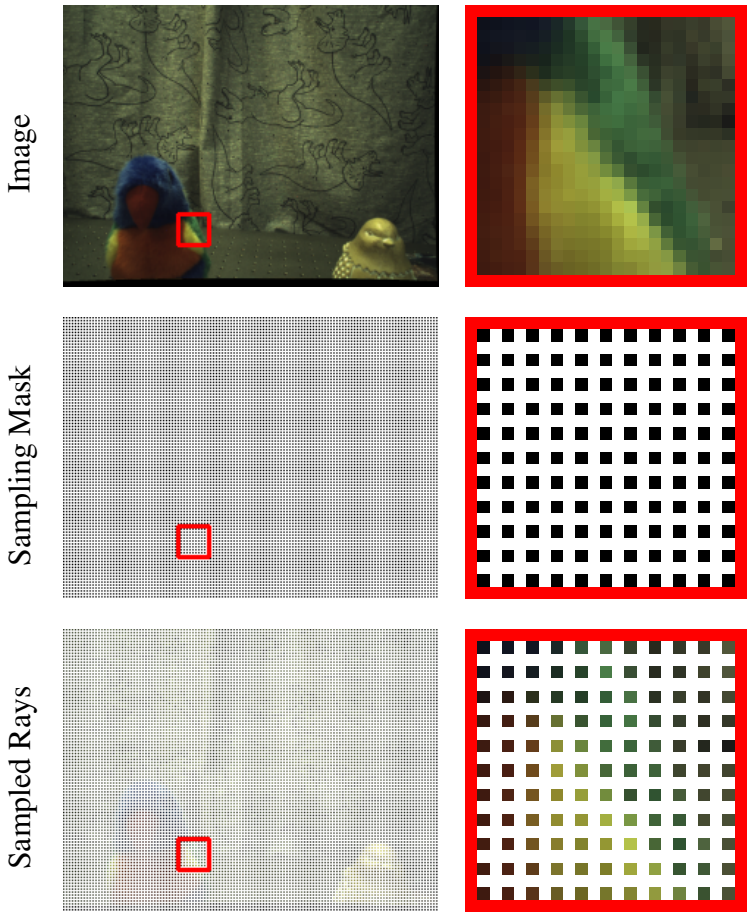

FIGURE 4.2. **Uniform sampling with sampling rate** $s = 0.25$**.** Top: the image before subsampling. Middle: the generated ray sampling mask, where black pixels represent a sampled ray. Bottom: the image after ray sampling. The final subsampled image has a structured uniform ray distribution across the image space.

subsampled set containing $s \cdot N$ rays. We can randomly select $s \cdot N$ rays by computing a random permutation of a sequence of N Boolean values, where $s \cdot N$ of the values are true, and the remaining $N - s \cdot N$ are false. Each Boolean value maps to a ray in the light field image. After shuffling this Boolean value sequence to compute a random permutation, the rays which correspond to a true Boolean value are then added to the subsampled set. This method ensures that exactly $s \cdot N$ rays are selected.

Alternatively, we can simply select each ray with a probability s, equivalent to performing N Bernoulli trials over the set of N rays. Performing this over all N rays produces a ray set size with an expected value of $s \cdot N$, matching the target sample size. Due to the each ray

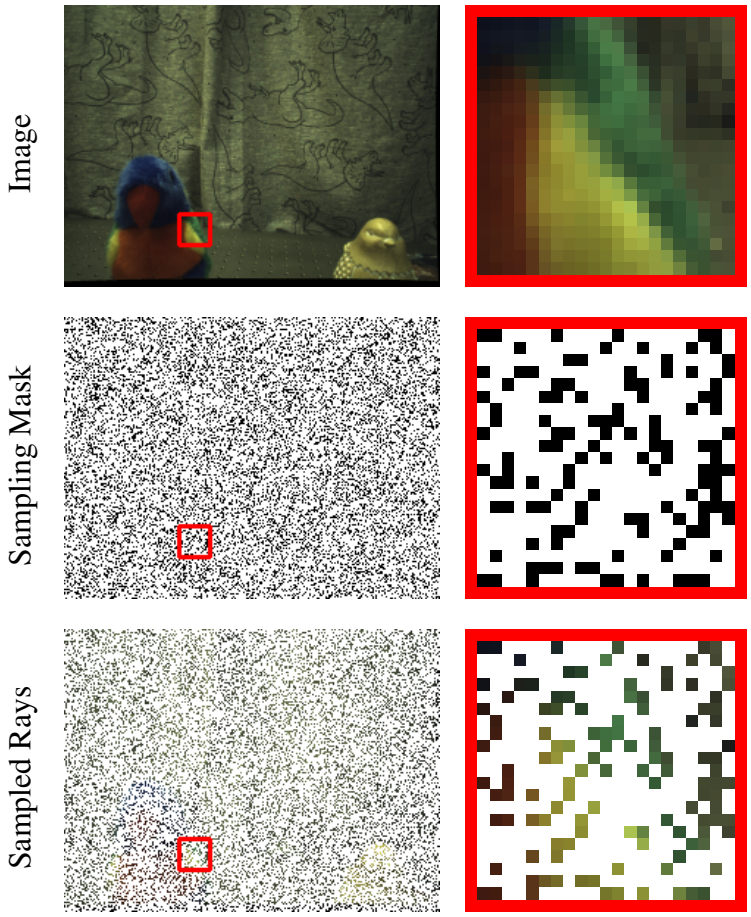

FIGURE 4.3. **Random sampling with sampling rate** $s = 0.25$**.** Top: the image before subsampling. Middle: the generated ray sampling mask, where black pixels represent a sampled ray. Bottom: the image after ray sampling. The distribution of rays is randomly scattered across the image space.

being selected as a separate random trial, the exact number of rays selected is not guaranteed. However, for large set sizes, only marginal differences from the target sample size are observed due to statistical convergence. From testing, we find no statistically significant difference between the performance of these two approaches. In our experiments, we utilise the second method for sampling, where every ray is sampled with a probability s.

4.2.3 View-Based Sampling

A key advantage of light field imaging over conventional imaging is that light field images contain angular information of the scene, as a single point in the scene may be observed

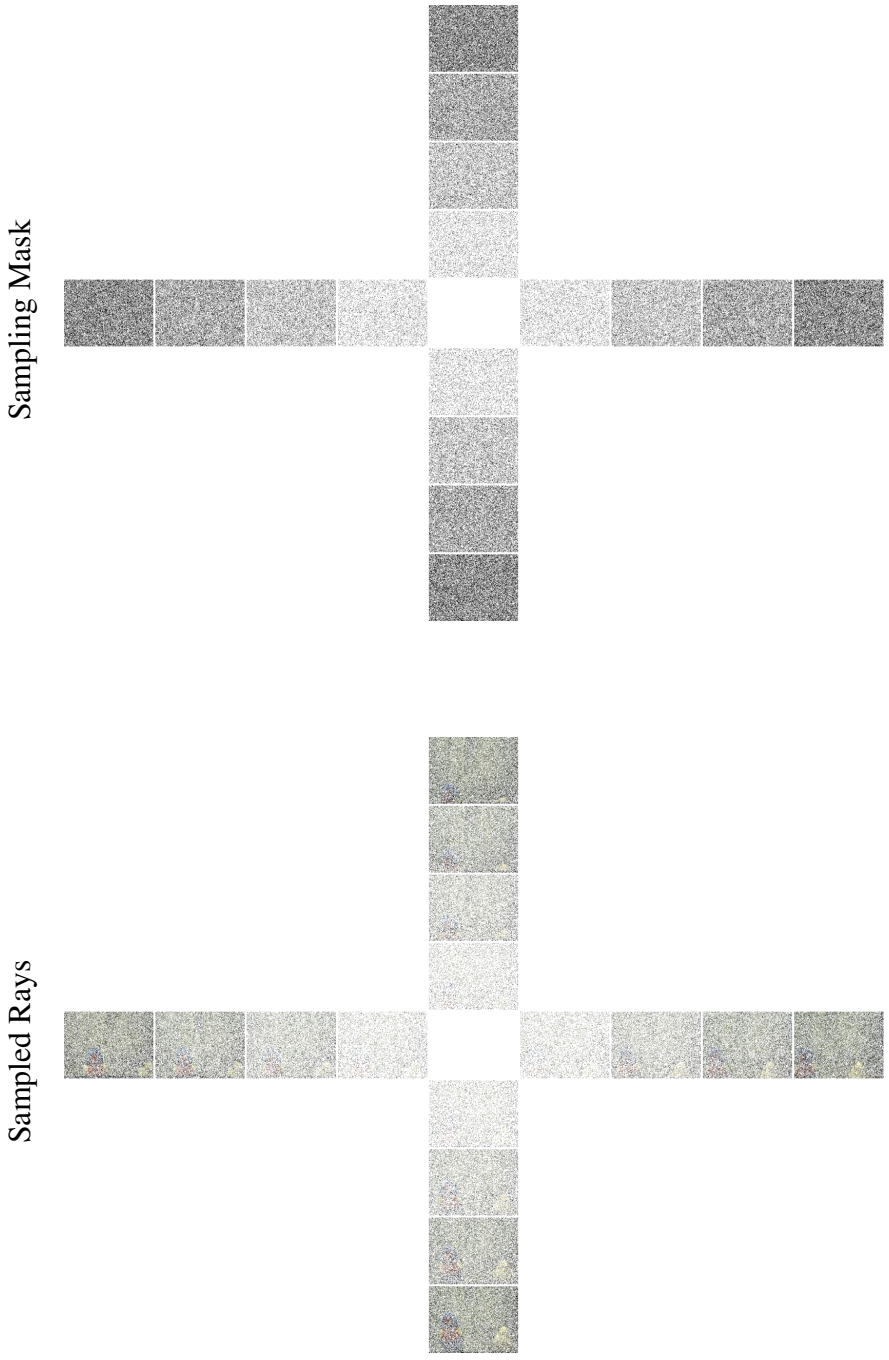

FIGURE 4.4. **View-based sampling with sampling rate** $s = 0.25$**.** Top: the generated ray sampling mask, where black pixels represent a sampled ray. Bottom: the image after ray sampling. The subsampled image has a ray distribution with higher densities at view positions which are further from the centre view.

multiple times by different subviews in the light field image. Typically, views of a scene that have camera poses closer to each other will capture similar information from the scene; this is what introduces redundant information in light field images. Views that are further apart should therefore generally capture more information of the scene.

To exploit this intuition, we develop a view-based sampling method which adjusts the sampling rate based on the position of the subview within the light field image. We theorise that by utilising a higher sampling rate for views with a larger camera baseline which are further from the centre light field subview, the subsampled rays will capture more useful information about the scene for performing view synthesis. In the results, we evaluate the correctness of this theory by comparing our pipeline performance using view-based sampling of different weighting distributions. Figure 4.4 shows an example ray subsampling of a light field image. In this example, view-based sampling is performed using a linear weighting based on distance: the sampling rate for a given subview is scaled proportionally depending on the distance from the centre subview.

The implementation of this sampling method requires a weighting w_i to be defined for each subview in the light field image, where w_i represents the weighting for the i-th subview, normalised such that $\sum_i^N w_i = 1$ (the sum of all weights is 1). From these defined weightings and the sampling rate s, we need to assign a sampling rate locally for each subview such that each local sampling rate s_i is proportional to its subview weighting w_i, while ensuring the global sampling rate across all subviews is equal to s. Since each subview contains the same number of rays, the global sampling rate s is equal to the average sampling rate across all of the subviews. Defining C as the number of subviews in the light field image, we attain the following condition.

$$\frac{1}{C}\sum_{i}^{N} s_i = s \tag{4.1}$$

Subsequently, we can derive the total sum of all local sampling rates across all of the subviews.

$$\sum_{i}^{N} s_i = C \cdot s \tag{4.2}$$

Each local sampling rate s_i can then be calculated by assigning a weighted portion of the total sampling rate sum, with the weighting determined by w_i.

$$s_i = C \cdot s \cdot w_i \tag{4.3}$$

We then perform random sampling separately on each subview using its local sampling rate to produce the final set of subsampled rays for training.

4.2.4 Image Gradient Sampling

The previous proposed sampling methods in this work do not take into account the measured information in each image, instead sampling on a specified distribution across the image space with no regard for the actual pixel values.

Intuitively, an optimal subsampling of a scene should account for the information contained in each image, as this allows the sampling method to exploit the observed information about the scene context, and produce a more informed decision on which rays to sample.

We hypothesise that to maximise view synthesis reconstruction quality, it is ideal to more densely sample regions of the scene which contain greater image gradients. The gradient of an image measures the directional intensity change in a local region. In an image, higher gradients signify greater changes in the intensity or colour in that local region, which typically correspond to image regions which contain object edges and changing textures. For this sampling method, we aim to use image gradients to sample more densely around these areas of interest to investigate whether this will facilitate higher quality reconstructions from the NeRF model.

As a preprocessing step, we first convert the image to greyscale, then apply a Gaussian blur to each image through convolution with a Gaussian kernel. Performing a blur operation reduces high frequency detail in the image, and Subsequently, we compute the image gradients for each subview of the light field image. Firstly, the image gradients are computed separately in the horizontal and vertical directions of the image using a Sobel filter. For both the horizontal and vertical directions, a convolution operation is performed on the image with a 3x3 kernel.

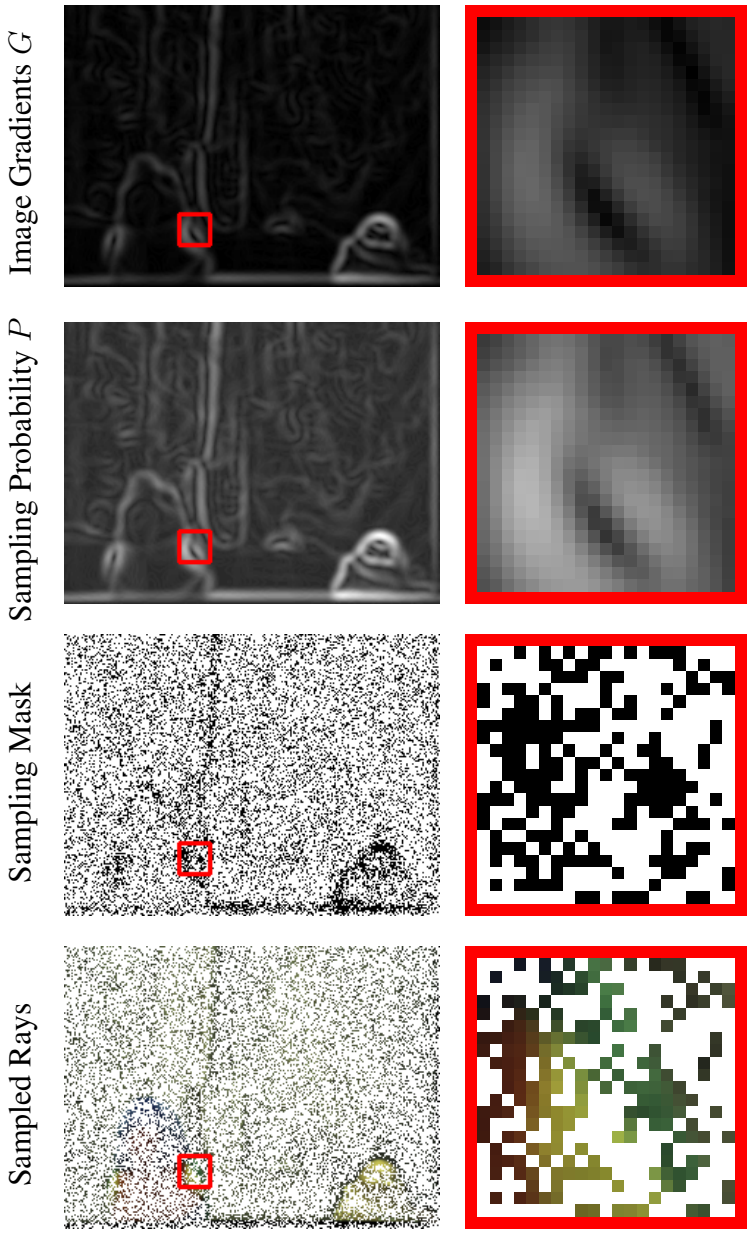

FIGURE 4.5. **Image gradient sampling with sampling rate** $s = 0.25$**.** Row 1: the image gradient magnitudes of the original image. Row 2: the sampling probability distribution P, where probabilities between 0 and 1 are mapped between black and white colours. Row 3: the generated sampling mask. Row 4: the image after ray sampling. The subsampled image has a ray distribution with higher densities at spatial locations with greater image gradients.

This produces the gradients in the horizontal direction G_x and the gradients in the vertical direction G_y. We then calculate the overall image gradient magnitude G using the horizontal and vertical directional image gradients, where $G = \sqrt{{G_x}^2 + {G_y}^2}$.

An example illustration of G is shown in the first row of Figure 4.5. This image gradient magnitude G defines the basis for our random sampling. We randomly sample on a per ray basis in a similar manner to the random sampling method, where each ray is selected with a certain probability. However, unlike random sampling, the probability distribution is not uniform; each ray does not have an equal selection probability. Instead, each ray should have a sampling probability proportional to the image gradient magnitude at the corresponding pixel. Therefore, a pixel with a higher image gradient should have an increased sampling probability. As we operate separately on each light field subview, we consider a subview with n pixels. For a sampling rate s, the target subsampled set size is $s \cdot n$. To assign a sampling probability p for each of the n pixels, two constraints must be satisfied:

(1) Each probability p must lie in the range $[0, 1]$. This constraint ensures that each p defines a valid probability.
(2) The sum of probabilities must equate to $s \cdot n$. The sum of all N probabilities is equivalent to the expected value of the subsampled set size. This constraint ensures that the expected number of sampled rays is equal to the target number of sampled rays for the given sampling rate.

The next step of this sampling method transforms the image gradient space G into a probability space P which satisfies the above constraints, where every probability value p in P contains denotes the probability of sampling a certain pixel. Every probability p is proportional to the corresponding image gradient value in G. To map the image gradients G to a range of probability values between 0 and 1 in order to satisfy the first constraint, we could construct P by simply normalising G through the division of each value by the maximum value in G, denoted by G_{max}. However, doing this does not guarantee the second constraint $\sum_{p \in P} p = s \cdot n$, causing the resultant subsampled set size to be incorrect. On the other hand, a P could be constructed to satisfy the second constraint by multiplying each value in G by a constant factor such that the total sum of probabilities is equal to the target value defined by the second constraint. However, doing this no longer guarantees the first constraint, as this scaling may scale probabilities to values above 1. Hence, it is apparent that is it impossible for a constant

scaling factor across all pixels to define a transformation from G to P while satisfying the specified constraints.

To solve this problem, we build off our first proposed method of transforming G to P, where every value in G is divided by G_{max}, the maximum value in G. This step scales all the values in G to within the range $[0, 1]$, satisfying the first constraint. We refer to this scaled space as G_{norm}. However, the second constraint is still not satisfied; we must consequently compute a transformation which allows the space to satisfy the second constraint, without breaking the first constraint. To do this, we scale G_{norm} based on the difference from 1 of each value g_{norm} in G_{norm}. By scaling each g_{norm} based on the difference from 1 instead of the g_{norm} value itself, we can define a transformation which ensures that the probabilities do not exceed 1. Therefore, for every g_{norm} value in G_{norm}, we scale up the probability by adding a term $k(1 - g_{\text{norm}})$, where $k \in [0, 1]$ is a constant value shared by all of the g_{norm} values in G_{norm}. Hence, the space G_{norm} undergoes the following transformation:

$$G_{\text{norm}} \rightarrow G_{\text{norm}} + k \cdot (1 - G_{\text{norm}}) \tag{4.4}$$

In the above equation, for the minimum value of $k = 0$, G_{norm} remains the same. For the maximum value of $k = 1$, G_{norm} becomes 1 for all $g_{\text{norm}} \in G_{\text{norm}}$. Any value of k in the range $[0, 1]$ scales the space between these two extremities. As any g_{norm} value in G_{norm} cannot exceed 1 if k is in the range $[0, 1]$, the first constraint cannot be violated from this transformation. Therefore, we can find a value of $k \in [0, 1]$ which scales the space to satisfy the second constraint, while also satisfying the first constraint. From the second constraint, which defines the sum of all probabilities to equal $s \cdot n$, we can produce the following equation:

$$\sum_{g_{\text{norm}} \in G_{\text{norm}}} (g_{\text{norm}} + k \cdot (1 - g_{\text{norm}})) = s \cdot n \tag{4.5}$$

This equation can be solved for k:

$$k = \frac{s \cdot n - \sum_{g_{\text{norm}} \in G_{\text{norm}}} g_{\text{norm}}}{\sum_{g_{\text{norm}} \in G_{\text{norm}}} (1 - g_{\text{norm}})} = \frac{s \cdot n - \text{sum}(G_{\text{norm}})}{\text{sum}(1 - G_{\text{norm}})} \tag{4.6}$$

Therefore, using this value for k allows both constraints to be satisfied, defining a valid probability space P. Therefore, we have solved for the transformation $G_{\text{norm}} \rightarrow P$.

$$P = G_{\text{norm}} + \frac{s \cdot n - \text{sum}(G_{\text{norm}})}{\text{sum}(1 - G_{\text{norm}})} \tag{4.7}$$

Using P, we can perform ray sampling on the image. We iterate through every probability p within the computed probability space P, and sample the corresponding ray with probability p. This achieves a sampling method where each ray is sampled at a probability proportional to the corresponding pixel image gradient, and the resulting subsampled set size has an expected value equal to the target size of $s \cdot n$. Figure 4.5 depicts each step of the sampling process.

4.2.5 Fixed Pattern Sampling Variations

We also develop variations on certain previously-established sampling methods. These variations introduce sampling in a fixed pattern across different source views of the scene. Sampling in a fixed pattern across views means that in each light field image taken from a different camera pose, rays from the same locations are always selected. With the previous sampling methods which utilise randomness, a random ray sampling is generated for each source view. However, in the fixed pattern variations, specific pixel positions are sampled once and then the generated sampling pattern is fixed and applied identically across all of the source views. The key motivation for this is to enable applications where low bandwidth is necessary, or where optimisation of computational memory usage is required.

Of the four ray sampling methods proposed earlier, we introduce two fixed pattern variations: fixed random sampling and fixed view-based sampling. These correspond respectively to the fixed pattern variations of random sampling and view-based sampling. Uniform sampling is already a fixed patterns sampling method, therefore there is no variation to be introduced. Image gradient sampling is a semantic sampling method, and therefore does not have a fixed pattern variation.

CHAPTER 5

Results

This chapter details the experiments and results used to evaluate our view synthesis pipeline and each of the proposed ray sampling methods.

We first outline the experimental setup, describing how the ideas and theories presented in the prior chapter have been realised into a practical implementation, and subsequently how the implementation is evaluated. After establishing the implementation and evaluation methods, we present the results of our evaluation process. Evaluation is split into multi-view and few-shot scenarios. In the multi-view scenario, our pipeline is evaluated with a larger number of source views similar to how conventional NeRF pipelines are trained. Subsequently, the few-shot scenario evaluates performance on an extremely challenging scenario where very few source views are available. We also evaluate optimisation of view-based sampling, and explore how different view weighting distributions affect view synthesis performance.

5.1 Experimental Setup

5.1.1 Light Field Camera

The input of our view synthesis pipeline requires a set of light field images taken from a light field camera. The light field camera used to collect the dataset for this work is the EPIModule sparse light field camera, which was previously discussed in the literature review of this work. This light field camera consists of a sequence of 17 conventional cameras arranged in a cross shape. Each conventional camera view captures a light field subview with a resolution of

256×192. Figure 5.1 shows the 17 subviews of an example image taken by the EPIModule camera.

FIGURE 5.1. **A light field image captured by the EPIModule sparse light field camera.**

5.1.2 Conventional Camera

For our experiments, we evaluate against a conventional imaging NeRF pipeline used as a performance baseline. To imitate a conventional camera imaging setup without requiring additional camera hardware, we simply extract the centre subview from each light field image to construct the equivalent conventional imaging dataset. However, this establishes a resolution imbalance between conventional and light field images for our datasets, as each light field image contains a factor of 17 more rays than the conventional image. In

FIGURE 5.2. **Example images from the dataset.**

our evaluations, we account for this imbalance by leveraging subsampling to evaluate both pipelines on the same number of trained rays.

5.1.3 Dataset

An existing dataset collected by Digumarti *et al.* [**1**] using the EPIModule light field camera was used for the evaluation of our pipeline. The dataset consists of sequences of light field images which capture a setup of various objects placed on a table. These images were collected by mounting the camera on a UR5e robot arm, which moved through the scene in a pre-planned trajectory, with images captured at a specified frequency. Some examples of images from the dataset are shown in Figure 5.2.

We distinguish between two different evaluation scenarios: multi-view and few-shot. For each of these scenarios, we further subset and split the data for training and evaluation. This will be further detailed in the respective sections for each evaluation scenario.

5.1.4 Determining Camera Parameters

In addition to the set of source images, our pipeline requires two additional inputs: the intrinsic and extrinsic camera parameters. The intrinsic camera parameters define the transformation

from each pixel on the camera sensor to the ray direction in space observed by that pixel. The extrinsic camera parameters refer to the camera poses. For each captured image, our pipeline requires the corresponding pose of the camera.

As a NeRF model trains on a set of rays, the camera parameters must be known to compute the transformation from pixels to rays. However, this transformation operates in the local reference frame for a single view. As NeRF requires all the rays to be defined in a global coordinate system, the camera poses with reference to a global coordinate system must be known. Thus, by knowing the camera pose for each source view, we can transform all the captured rays into a global frame upon which a NeRF model can operate.

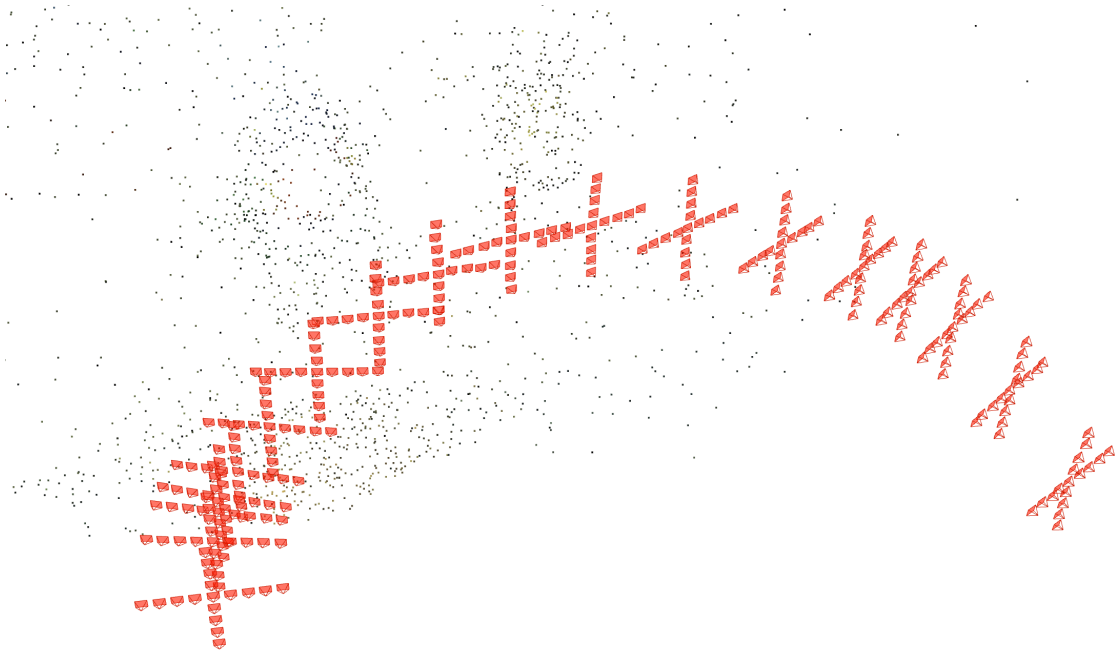

FIGURE 5.3. **COLMAP reconstruction example.** Each red frustum represents a conventional image view, with each cross-shaped arrangement of frustums representing a full light field view from the sparse light field camera. Each of the coloured points represents a single feature in the scene which was extracted and matched across views during the COLMAP reconstruction.

To estimate camera intrinsics and poses, we use the COLMAP [**23**] structure from motion (SfM) pipeline. SfM algorithms reconstruct three-dimensional scene geometry from sequences of two-dimensional images. COLMAP extracts features from each image, and matches features across images to incrementally reconstruct the scene geometry. The pipeline also

simultaneously solves for the intrinsic and extrinsic camera parameters for all of the images, which we use as input to our view synthesis pipeline. Figure 5.3 shows an example output of the reconstructed features and camera poses from the COLMAP pipeline.

However, computing camera parameters and poses using SfM algorithms has certain limitations, and will run into inaccurate estimations and failures in extremely challenging contexts. Such scenarios may involve extreme low light conditions, scenes with high specularity and reflective surfaces, or occlusions such as snow or fog. In this work, we are able to rely on COLMAP to produce accurate estimations of camera intrinsics and extrinsics as the scenes in our dataset do not exceed the performance limits of COLMAP. For future work which may address these more challenging contexts where SfM algorithms fail, collecting ground truth camera poses and parameters may be necessary for evaluation.

5.1.5 NeRF Implementation

Many implementations and variations of NeRF exist, each offering various advantages. A majority of implementations, including the original NeRF paper [**3**], are written in Python using various machine learning libraries and frameworks. These typically run quite slow, sometimes taking in the range of hours up to days to converge for larger inputs.

We selected NVIDIA's InstantNGP [**2**] implementation of NeRF for our pipeline as it is both computationally efficient and highly accessible. InstantNGP's typical NeRF convergence times are in the range of seconds to minutes, due to its implementation in C++ and the addition of multi-resolution hash encoding.

5.1.6 Evaluation

For each evaluation scenario, we reserve some of the image data for evaluation such that the pipeline does not train on the evaluation data. A single sample from the evaluation dataset comprises a ground truth image and the corresponding camera pose. Once the NeRF model in our pipeline is trained, we query the target camera pose from the evaluation data to the

trained model to reconstruct a rendering of the view seen by that camera. Subsequently, we can compare this rendered image with the captured ground truth image of the scene.

To quantitatively compare between the reconstruction and the ground truth, we employ two metrics: peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). PSNR

These metrics are an adopted standard used widely across the research literature in novel view synthesis. Both metrics quantify the similarity between two images. PSNR, which is measured in decibels (dB), measures the ratio of meaningful information in the image relative to the noise present. Hence, a higher PSNR value indicates less noise present in the image, which relates to higher reconstruction quality.

SSIM is a more recent evaluation metric which defines a measure based on the luminance, contrast, and structure of an image. This metric is designed to more appropriately model image similarity as perceived by the human visual system in comparison to PSNR, and therefore is considered a more accurate depiction of reconstruction quality for a synthesised view [**24**]. Therefore, while both metrics are considered, we attribute more value to SSIM. SSIM scores reside between 0 and 1, with higher values indicating higher similarity between images.

NeRF trains in iterations, where each iteration involves performing gradient descent on a single batch of rays in the training data. While training our pipeline, we evaluate performance on the above metrics every 100 iterations, and train until the training loss has approximately converged.

5.2 Multi-View Performance

We first perform evaluation on the multi-view scenario, where there are a sufficient number of source views provided such that the conventional imaging NeRF pipeline is able to converge to a reasonable solution. Later, we explore the few-shot scenario, which presents a much greater challenge for the conventional NeRF pipeline. For the multi-view case, we train each

pipeline on six source views: six conventional images for the conventional imaging pipeline, and six light field images for our light field pipeline. Evaluation is performed on a single target camera pose. Figure 5.4 shows the six source views which are used for training and the single ground truth target view.

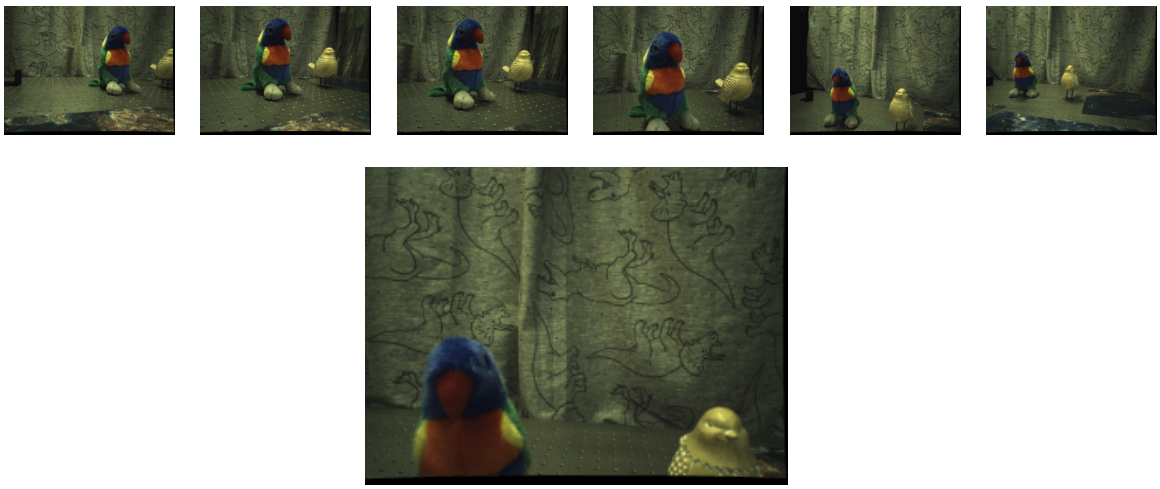

FIGURE 5.4. **An example dataset for the multi-view scenario.** Top: the six source views used for training. Bottom: the ground truth target view used for evaluation.

5.2.1 Reconstruction Quality

For the imaging hardware used in our experimental setup, each light field image contains a factor of 17 more data than the conventional camera image. Hence, training on full light field images establishes an unfair comparison with the conventional imaging pipeline. To ensure a fair comparison between different methods, we train all methods using the same number of measured light rays of the scene. For the conventional imaging pipeline, the full set of rays within the conventional image is used for training. For our light field pipeline, we downsample the full set of rays by a factor of 17 using each of the ray sampling approaches with $s = \frac{1}{17}$. This ensures that view synthesis is performed on a consistent number of ray measurements of the scene across different pipelines. To establish a performance standard for comparison, we also show reconstruction results for training on the full number of rays for each light field image.

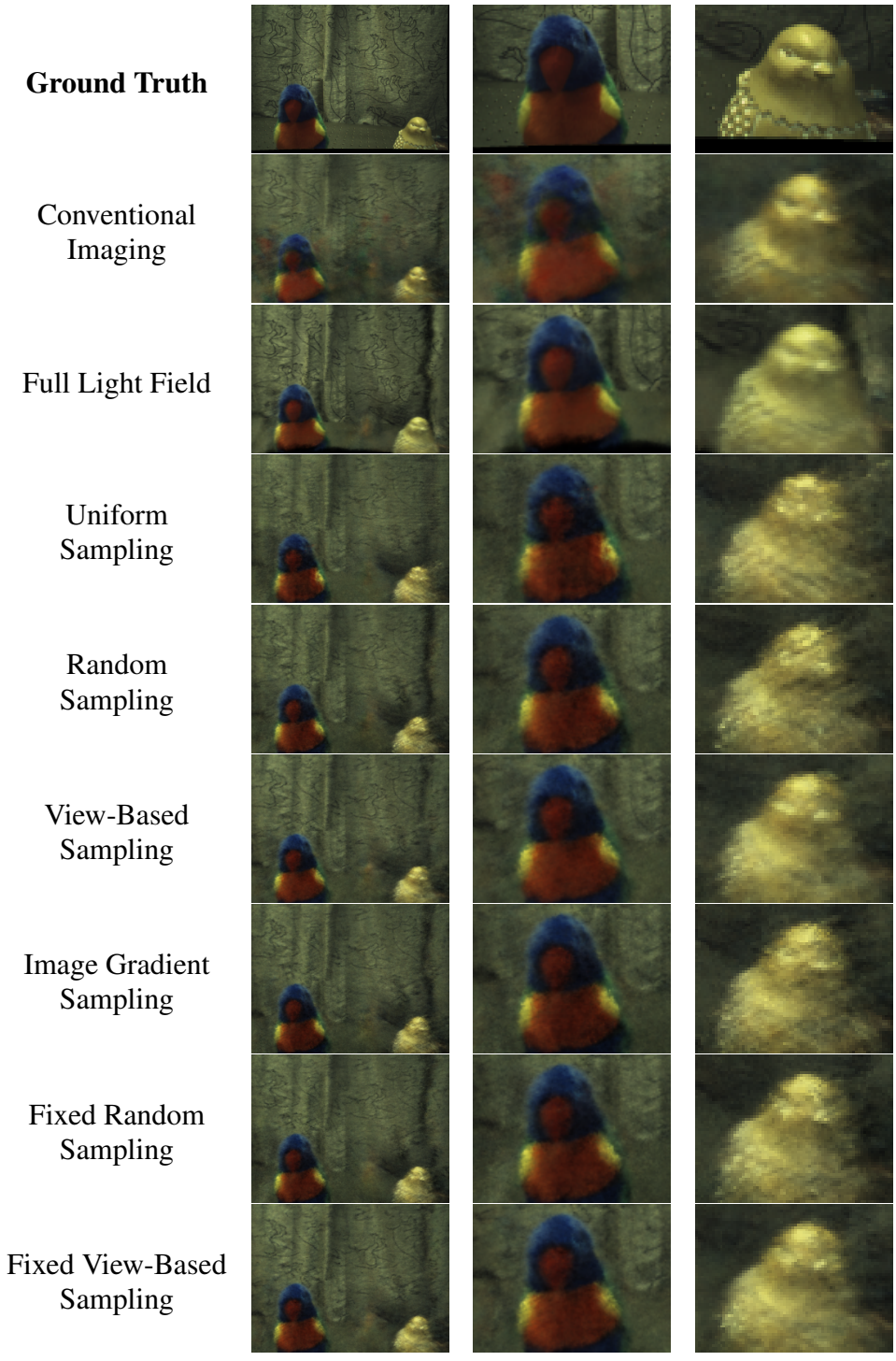

FIGURE 5.5. **Multi-view scenario: reconstructed views compared against the ground truth.** The conventional imaging reconstruction contains evident ghosting artifacts and significant blur around the left bird, which are not present in the light field methods. However, the right bird is reconstructed better by the conventional imaging pipeline than the subsampled light field pipelines.

Figure 5.5 shows the reconstructed views of each method in comparison to the ground truth target view. The first column of images show the full reconstructed image, while the next two columns illustrate zoomed sections of the image. The conventional imaging pipeline performs reasonably well, with the reconstructed view appearing visually similar to the ground truth image. The left bird reconstruction is worse than all other methods: the beak is faded and not distinct, and much more blur is present. Additionally, there are ghosting artifacts surrounding the left bird, which are not present in any of light field method reconstructions. However, the right bird is reconstructed well in comparison to other subsampled light field methods. The beak and eye of the right bird are distinguishable; many of the other reconstructions fail to clearly reconstruct this feature.

As expected, the reconstruction using the full light field produces the highest quality reconstruction due to being trained on significantly more input data. The left bird is reconstructed accurately, and a considerable amount of the textures on the background curtain have been reconstructed. The texture on the body of the right bird is also identifiable, though quite blurry.

The uniform and random sampling reconstructions are visually comparable. Both methods perform poorly on the right bird; a significant amount of detail present in the ground truth image is lost in the reconstruction. The remaining sampling approaches all perform similarly, reconstructing the scene fairly accurately with the correct appearance and geometry.

5.2.2 Convergence

Next, to quantitatively evaluate the reconstructions for each method, we examine the PSNR and SSIM scores for the synthesised views produced by each method. Instead of solely examining the metric scores on reconstructions produced by the converged NeRF model, we periodically evaluate metrics scores every 100 iterations during training. This allows us to compare and evaluate the rate of convergence for different methods.

Figure 5.6 shows the convergence of PSNR and SSIM scores across the training process. Image gradient sampling achieves the highest score in PSNR performance, but performs

poorly on SSIM. Full light field reconstruction dominates on SSIM, but falls short on PSNR. By visually comparing to the reconstruction quality observed in Figure 5.5, we see that SSIM is a significantly better indicator of visual reconstruction quality. While the full light field method achieves the best visual result as seen in Figure 5.5, it ranks last among all methods in converged PSNR scores. Hence, the PSNR results are evidently not representative of the reconstruction quality produced in this experiment. From the converged SSIM scores, we see that the view-based sampling approaches outperform all other methods which use the same amount of input data for training. Conventional imaging performs worse than the view-based sampling methods, but outperforms the other subsampled light field methods. Table 5.1 contains the converged numerical metric values for each of the sampling methods.

In addition to the converged metrics scores, we also evaluate the rate of convergence across training. While conventional imaging performed well on converged metrics scores, the rate of convergence is poor. In both the PSNR and SSIM plots in Figure 5.6, the performance lags significantly behind the light field methods in the first half of the training process, though eventually it reaches and surpasses the performance of all subsampled light field methods in SSIM, excluding the view-based sampling methods. Overall, the convergence results show a favourable advantage to methods utilising light field images. In this experiment, the light field methods reach approximate convergence in SSIM with almost half the number of training iterations of the conventional imaging pipeline.

TABLE 5.1. **Multi-view scenario: reconstruction performance for each sampling method.** View-based sampling outperforms all other methods on SSIM. Conventional imaging achieves the second highest SSIM score, slightly outperforming many of the light field methods.

Sampling Method	Number of Rays	PSNR	SSIM
Full Light Field	5,013,504	22.74	0.7406
Conventional Imaging	294,912	23.14	0.6351
Uniform	294,912	23.16	0.6123
Random	294,912	23.24	0.6235
View-Based	294,912	22.96	0.6526
Image Gradient	294,912	**23.46**	0.6142
Fixed Random	294,912	23.22	0.6210
Fixed View-Based	294,912	22.99	**0.6532**

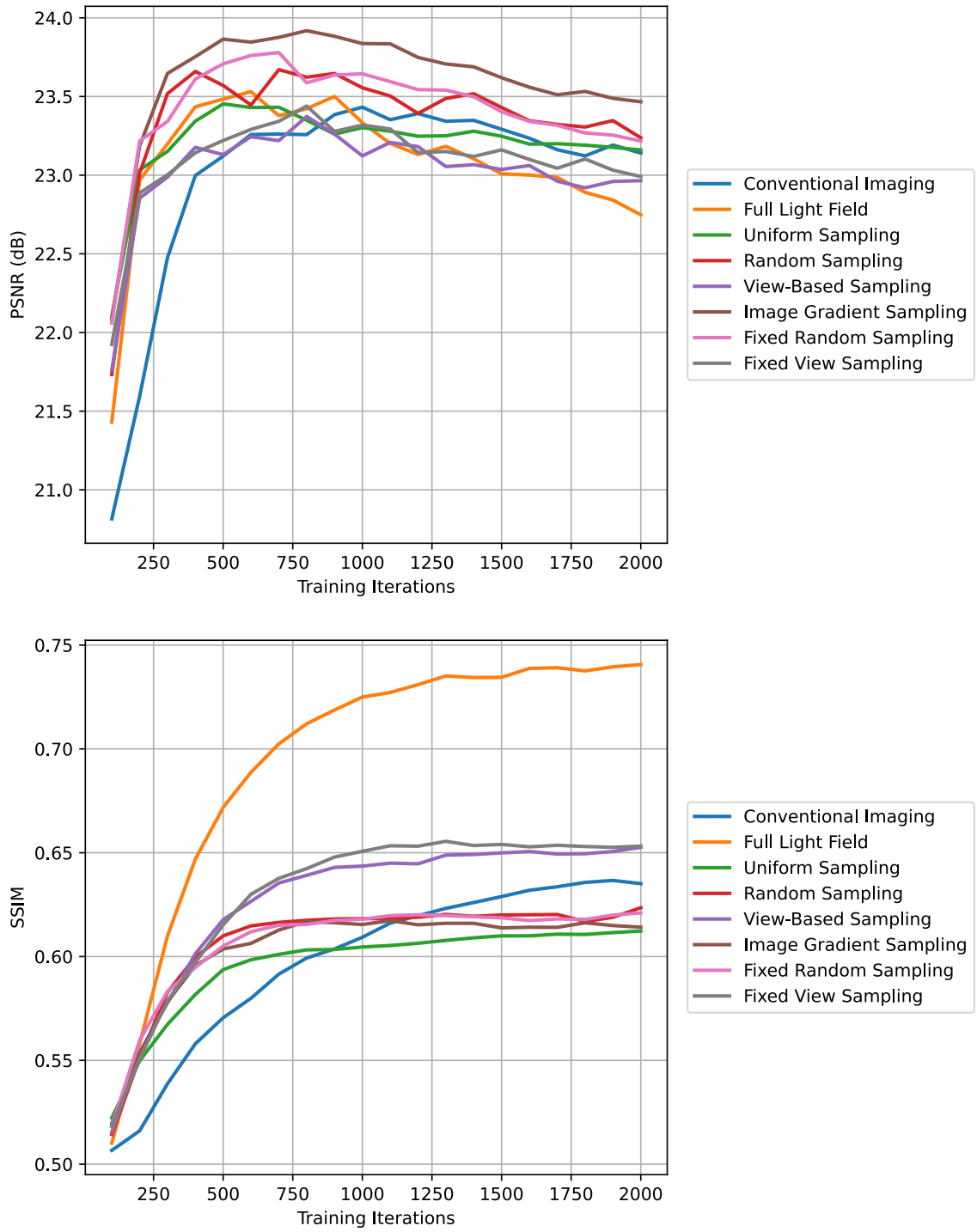

FIGURE 5.6. **Multi-view scenario: reconstruction performance across the training process for $s = 1/17$.** On both metrics, conventional imaging converges slower than all light field methods. However, the converged performance of conventional imaging on SSIM outperforms most subsampled light field methods, only falling short of view-based sampling.

5.2.3 Data Fidelity

In the above results, we evaluated on a fixed sampling rate (i.e. a fixed number of rays) for each of the ray sampling methods. While certain ray sampling methods may perform better than others for one sampling rate, this does not necessarily guarantee that the same method will perform better for all sampling rates. Therefore, to concretely determine which sampling methods are better, we evaluate the converged performance in reconstruction quality across many different sampling rates. A ray sampling method which consistently achieves higher reconstruction quality over another method for all sampling rates can consequently be considered a more optimal sampling approach.

This evaluation qualitatively measures the data fidelity for different sampling approaches. Data fidelity expresses how accurately data characterises the source. In the context of ray sampling, data fidelity expresses how accurately a subsampled set of rays characterises the scene. If one set of rays results in a higher reconstruction quality in comparison to another set, it can be said to have higher data fidelity as it better characterises the appearance and geometry of the imaged scene.

Figure 5.7 shows the data fidelity plot for both PSNR and SSIM metrics. This is generated by evaluating the converged performance of each ray sampling approach over various sampling rates. The horizontal axis depicts the number of rays used for training, which is directly proportional to the sampling rate for the subsampled light field methods. We represent the horizontal axis as numbers of rays instead of sampling rates since sampling rates are only defined for the subsampled light field methods which utilise ray sampling. The conventional imaging and full light field approaches do not use sampling and hence do not have a defined sampling rate, though both train on a constant number of rays. As the number of rays approaches either zero or the full number of set of light field rays, the performance of each subsampling approach should converge towards each other; this is observed in the data fidelity plots. This occurs because the input data becomes more similar at both of these extremities. When the number of rays becomes zero, there is no input data to train on, therefore any approach should achieve the same performance. When the number of rays reaches the full

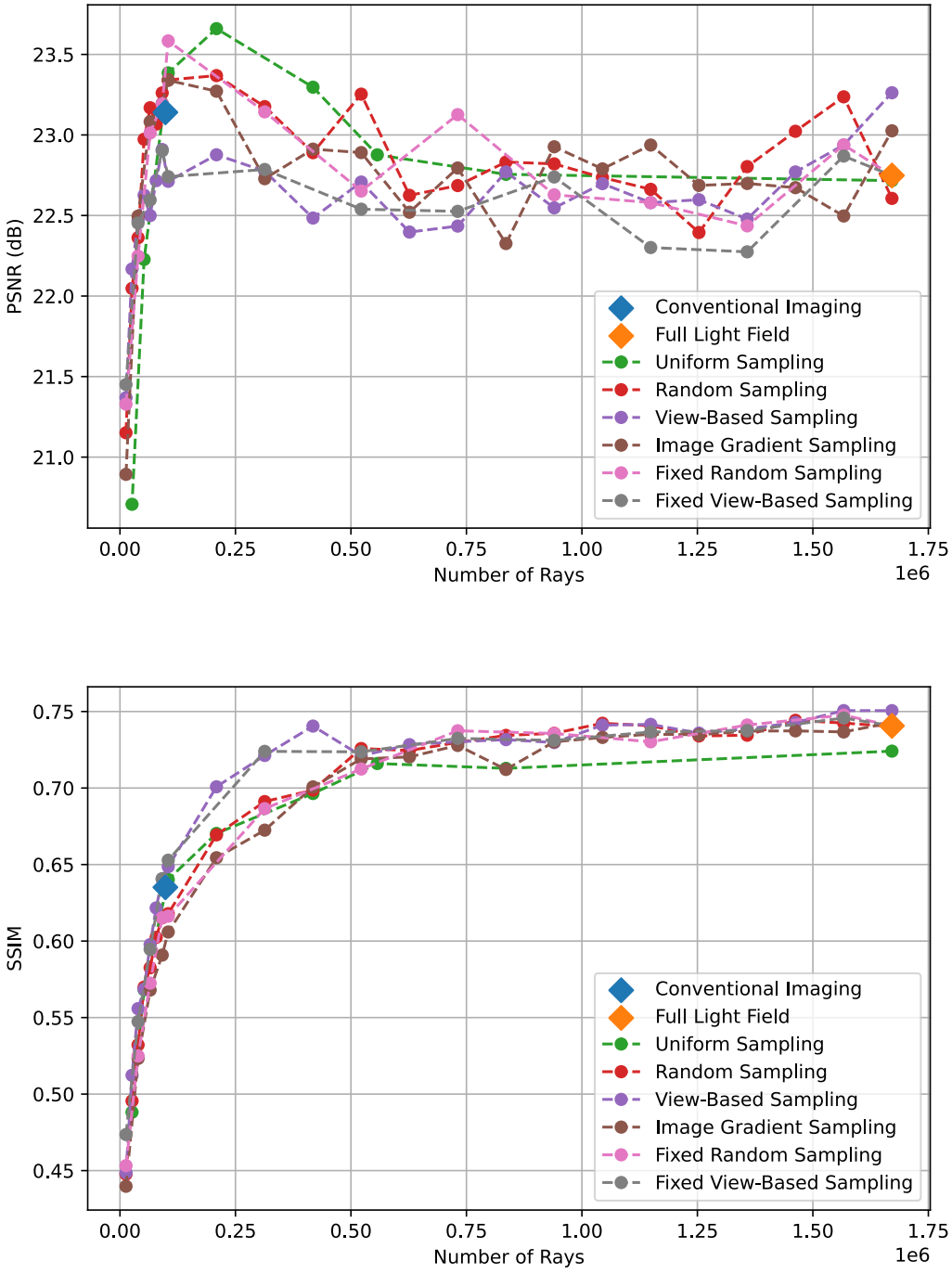

FIGURE 5.7. **Multi-view scenario: data fidelity evaluation.**. In the SSIM plot (bottom), view-based sampling shows a greater area under the curve, demonstrating a significant increase in reconstruction quality over other methods. All other subsampling methods perform comparably to each other.

number of rays in the light field, all the measured data is being trained on, as the sampling rate becomes $s = 1$. This case is equivalent to the full light field approach, where no subsampling is performed. All the numbers of rays between these two extremities are the values of interest for evaluation. A better subsampling approach will demonstrate higher reconstruction performance on these values, and thus achieve a larger area under the data fidelity curve.

From the SSIM data fidelity plots in Figure 5.7, view-based sampling demonstrates vastly superior performance to the other methods. The other subsampling methods achieve very similar performance. Image gradient sampling performs slightly worse than all other sampling methods for almost all of the evaluated ray numbers. Additionally, each of the fixed pattern variations perform on par with their variable pattern counterparts.

5.2.4 Summary

Overall, for the multi-view scenario, we have demonstrated considerable advantages of our light field imaging pipeline over the conventional imaging pipeline. Each of the light field methods demonstrate significantly faster training convergence. We found view-based sampling achieved the highest reconstruction quality of all the sampling methods, surpassing conventional imaging in both rate of convergence and converged performance. However, while conventional imaging demonstrated vastly inferior training convergence, it was able to achieve slightly better converged performance in comparison to many of the light field methods in this evaluation scenario.

5.3 Few-Shot Performance

Following the evaluation on the multi-view scenario, we explore the few-shot scenario, which poses a significantly harder challenge for view synthesis techniques. In this evaluation scenario, only two views of the scene are available. These two views are shown in Figure 5.8, along with the ground truth target view used for evaluation. In this section, we mirror the structure of the multi-view evaluation, examining the visual reconstruction quality, rate of convergence, converged metric performances, and data fidelity.

FIGURE 5.8. **An example dataset for the few-shot scenario.** Top: the two source views used for training. Bottom: the ground truth target view used for evaluation.

5.3.1 Reconstruction Quality

Figure 5.9 illustrates the reconstructed views for each of the sampling approaches. The reconstruction result from the conventional imaging pipeline is geometrically incoherent: the textures of the background curtain cannot be identified, the first bird (on the left) is significantly blurred, and the second bird (on the right) is entirely unrecognisable as a distinct object in the scene. Evidently, the model has been unable to learn the structure of the scene from the given input data.

However, the full light field reconstruction converges to a drastically better result. The detailed textures on the background curtain are clearly reconstructed and the first bird is almost visually identical to the ground truth. While the second bird is distinguishable in the reconstruction, very little of the texture and detail has been reconstructed. The second bird is particularly challenging to reconstruct in this scenario due to only half the bird being captured in the first source view, as shown in Figure 5.8. Additionally, given the further distance of the bird from the camera and the relatively low resolution of the source images, there are significantly fewer measured rays of the second bird compared to the multi-view scenario. However, we again emphasise the unfairness of comparing the full light field approach with conventional imaging

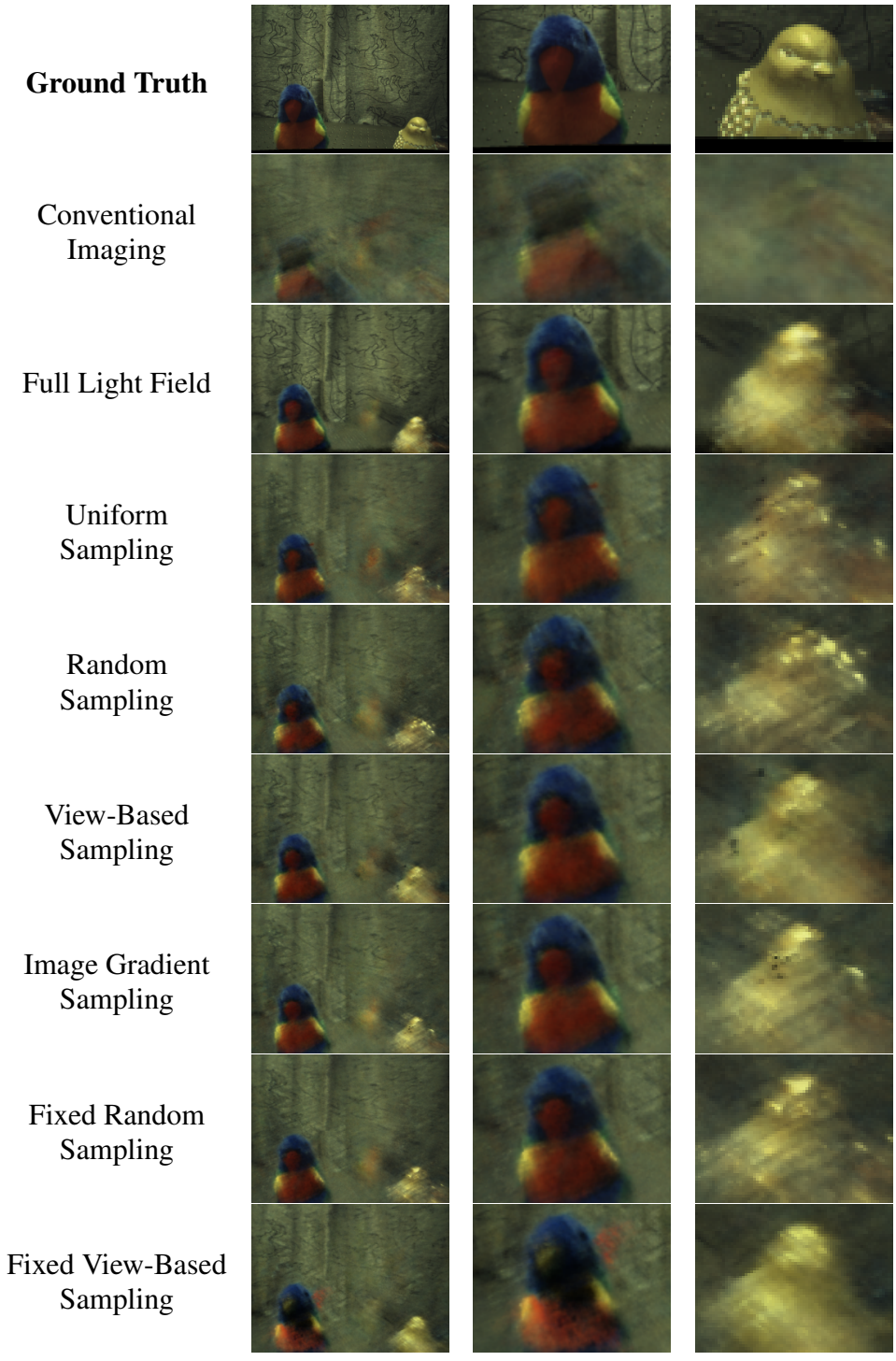

FIGURE 5.9. **Few-shot scenario: reconstructed views compared against the ground truth.** The conventional imaging pipeline fails to learn an accurate scene representation, producing a geometrically incoherent reconstruction. Contrarily, each of the light field approaches converge to an accurate representation of the scene geometry and appearance.

due to the input data difference: each light field view contains 17 times more data than the conventional image view.

A fair comparison can be evaluated between the conventional imaging pipeline and the subsampled light field methods, as in this evaluation scenario they are all trained on the same amount of input data. Interestingly, even with a drastic downsampling rate of $s = \frac{1}{17}$, the reconstructions from the subsampled approaches do not appear significantly worse than the full light field method. Every subsampled light field approach converges to a geometrically coherent scene representation. The left bird is accurately reconstructed, closely resembling the ground truth. However, a considerable amount of detail is lost in the reconstructions: while the right bird is distinguishable as a distinct scene object, very little of the details are captured. Similarly, a majority of the textures and details of the background curtain are missing. Additionally, a ghosting artifact between the two birds is present in all reconstructed views. This artifact corresponds to a location in the scene which is not observed by both of the source views. Hence, the scene representation is unable to accurately determine the scene appearance at that location. Ultimately, a considerable loss in quality is expected from the ground truth, as this evaluation scenario is extremely challenging. Notably, the results from this scenario highlight the drastic quality difference between conventional image and light field pipelines.

5.3.2 Convergence

We subsequently analyse the training convergence plots, shown in Figure 5.10. The convergence results for the conventional imaging pipeline reaffirm the poor conventional imaging reconstruction result from Figure 5.8. On the SSIM plot from Figure 5.10, the conventional imaging performance barely changes across the training process. From the input training data, the model is unable to converge to an accurate scene representation. Among the subsampled light field approaches, the results reflect the multi-view scenario, where view-based sampling again achieves the highest performance in SSIM. The other sampling approaches again perform similarly. Table 5.2 shows the numerical metrics values for the converged performance of each sampling method.

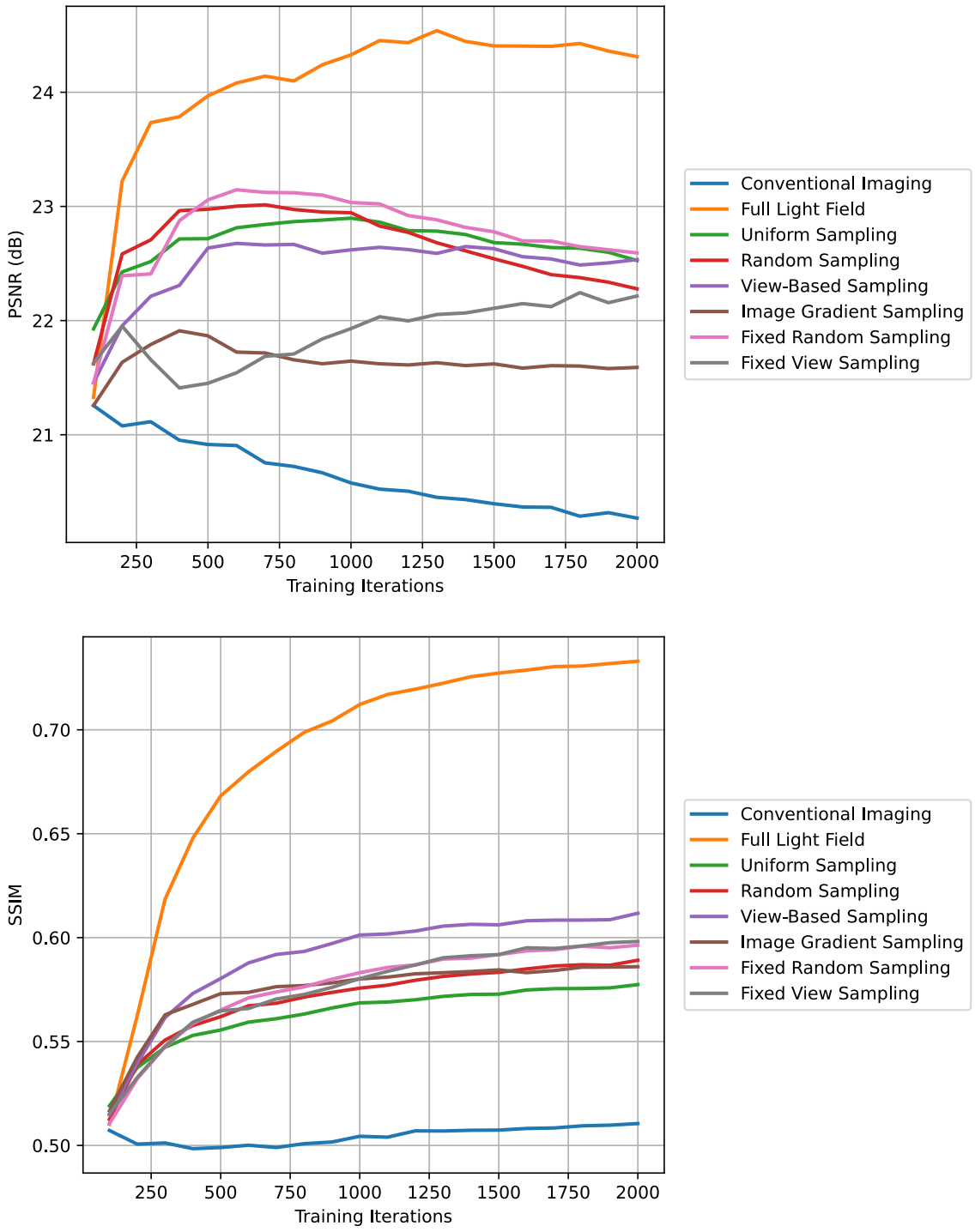

FIGURE 5.10. **Few-shot reconstruction performance across the training process for $s = 1/17$.** The conventional imaging pipeline fails to improve in performance on both metrics through the training process. The light field approaches are able to converge to significantly better results. On the SSIM metric (bottom), view-based sampling achieves the best performance of all subsampled light field approaches.

TABLE 5.2. **Few-shot scenario: reconstruction performance for each sampling method.** The conventional imaging pipeline performs drastically worse than all of the subsampled light field methods for both PSNR and SSIM. View-based sampling achieves considerably higher performance on SSIM than other methods.

Sampling Method	Number of Rays	PSNR	SSIM
Full Light Field	1,671,168	24.31	0.7330
Conventional Imaging	98,304	20.27	0.5105
Uniform	98,304	22.53	0.5774
Random	98,304	22.28	0.5891
View-Based	98,304	22.53	**0.6117**
Image Gradient	98,304	21.59	0.5860
Fixed Random	98,304	**22.59**	0.5964
Fixed View-Based	98,304	22.22	0.5982

5.3.3 Data Fidelity

Finally, we examine data fidelity, evaluating reconstruction performance of different sampling approaches across a range of sampling rates. The results of this are shown in Figure 5.11. In accordance with the multi-view evaluation, view-based sampling again demonstrates significantly higher SSIM reconstruction performance across all ray numbers. By examining the vertical slice intersecting the conventional imaging data point (shown in blue), we see the significant performance gain of the light field methods when trained on the same amount of data. By examining the horizontal slice intersecting the same data point, we quantify how much data the light field methods require to achieve the same reconstruction performance as conventional imaging pipeline. The SSIM plot indicates that the light field methods surpass conventional imaging performance using approximately *a factor of three less data*. Again, in this scenario, the fixed pattern variations achieve equivalent performance to their variable pattern versions.

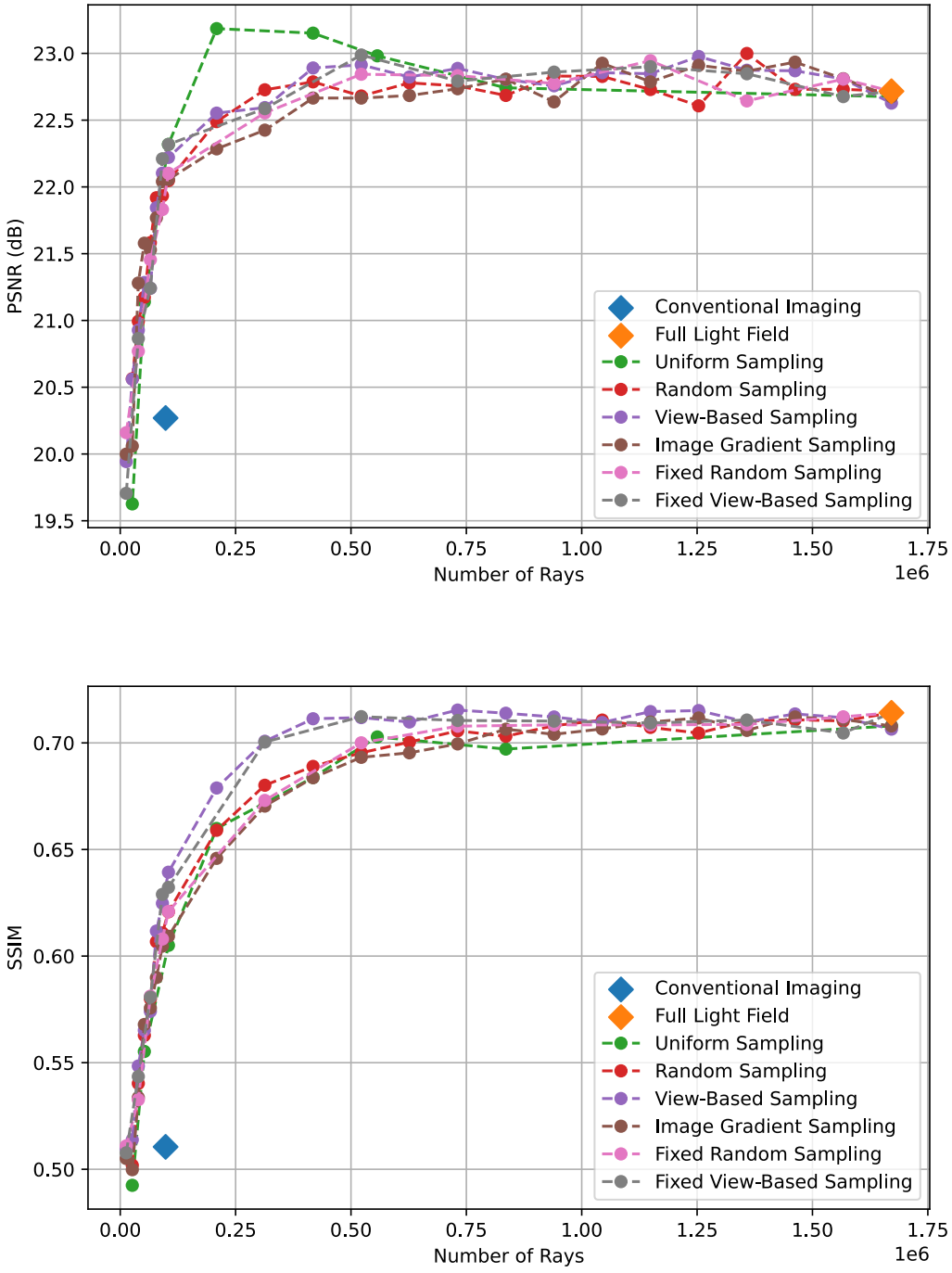

FIGURE 5.11. **Few-shot scenario: data fidelity evaluation.** From the SSIM plot (bottom), view-based sampling achieves the highest performance across all sampling rates. By examining the horizontal slice intersecting the conventional imaging data point (blue), we see that the light field methods surpass conventional imaging performance with *a factor of three less data*.

5.3.4 Summary

In the multi-view scenario, the light field methods demonstrated slight performance advantages over the conventional imaging pipeline. In this section, evaluating on the more difficult few-shot scenario unveiled drastic performance differences across the two pipelines, with light field methods capable of achieving significantly higher performance while simultaneously leveraging less training data. Across the subsampling approaches, view-based sampling still outperforms every other approach in reconstruction quality.

5.4 View-Based Sampling Optimisation

The results presented in this work so far have demonstrated view-based sampling to perform considerably better than the other proposed sampling methods. In this section, we investigate further improving view-based sampling performance by comparing various view weighting distributions. Initially, we proposed a linear distance-weighted weight distribution, where the sampling rate of each subview was proportional to the distance from the centre subview. This was illustrated in the sampling mask in Figure 4.4.

For the purposes of this analysis, we introduce the concept of the view distance d_v, which quantifies a distance from the centre subview. The subviews which are directly adjacent to the centre subview are assigned a view distance $d_v = 1$. Successive subviews going outwards from the centre subview are assigned discretely increasing view distances. For the EPIModule sparse light field camera used in our experimental setup, the 17 subviews occupy five discrete view distances $d_v = 0, 1, 2, 3, 4$. $d_v = 0$ corresponds to the centre subview. The remaining 16 subviews are uniformly distributed across the remaining four view distances $d_v = 1, 2, 3, 4$.

We investigate the impact of sampling rays at different view distances while maintaining a constant number of sampled rays. We compare four view-based sampling approaches, where rays are only sampled from one discrete view distance from $d_v = 1, 2, 3, 4$. As each of these view distances correspond to four subviews, for a given view distance we simply train on all the subviews at that view distance. Hence, the global sampling rate across each light field

image is a constant $s = \frac{4}{17}$, comprising four subviews where the local sampling rate is $s_i = 1$, and every other subview is not sampled.

We perform evaluation using our few-shot dataset, as this more challenging scenario is able to better distinguish differences in reconstruction performance.

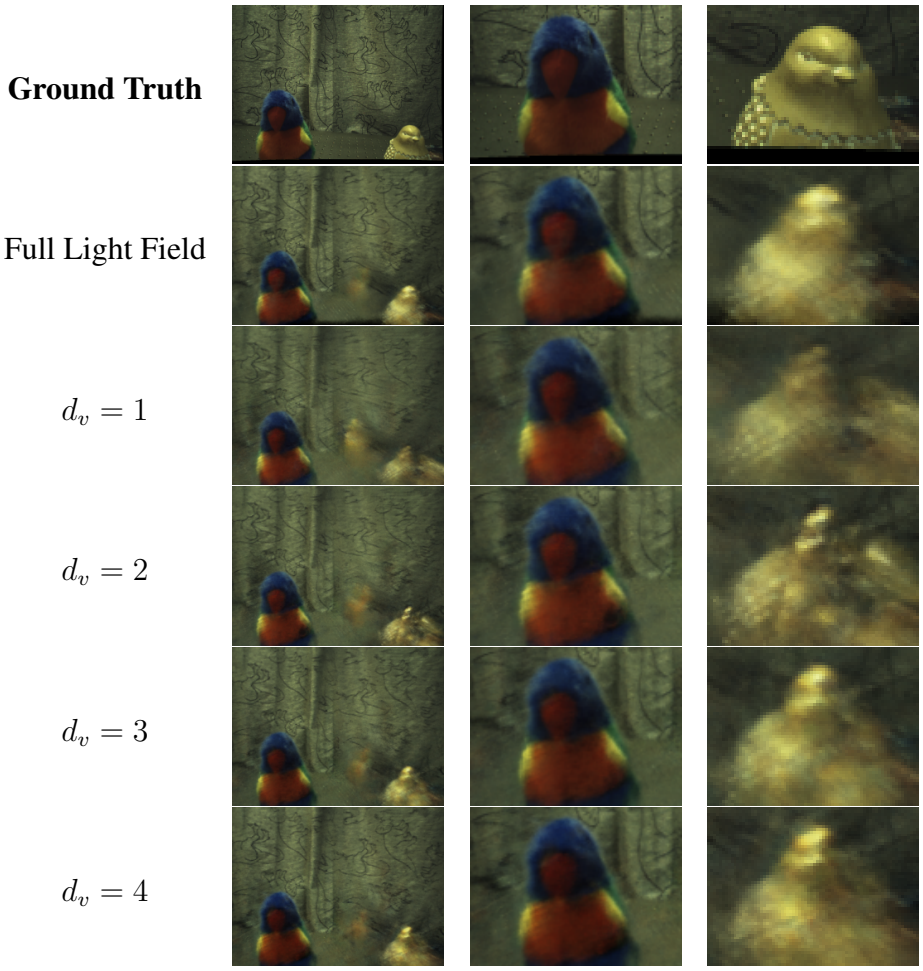

FIGURE 5.12. **Reconstructed views for view-based sampling with different view distances.** The right bird is reconstructed with higher visual quality for larger view distance values. Furthermore, the ghosting artifact between the two birds is less prominent for larger view distances.

5.5 Reconstruction Quality

Figure 5.12 shows the reconstructed views for view-based sampling with different view distances. While reconstructions of the left bird are visually similar and differences in quality

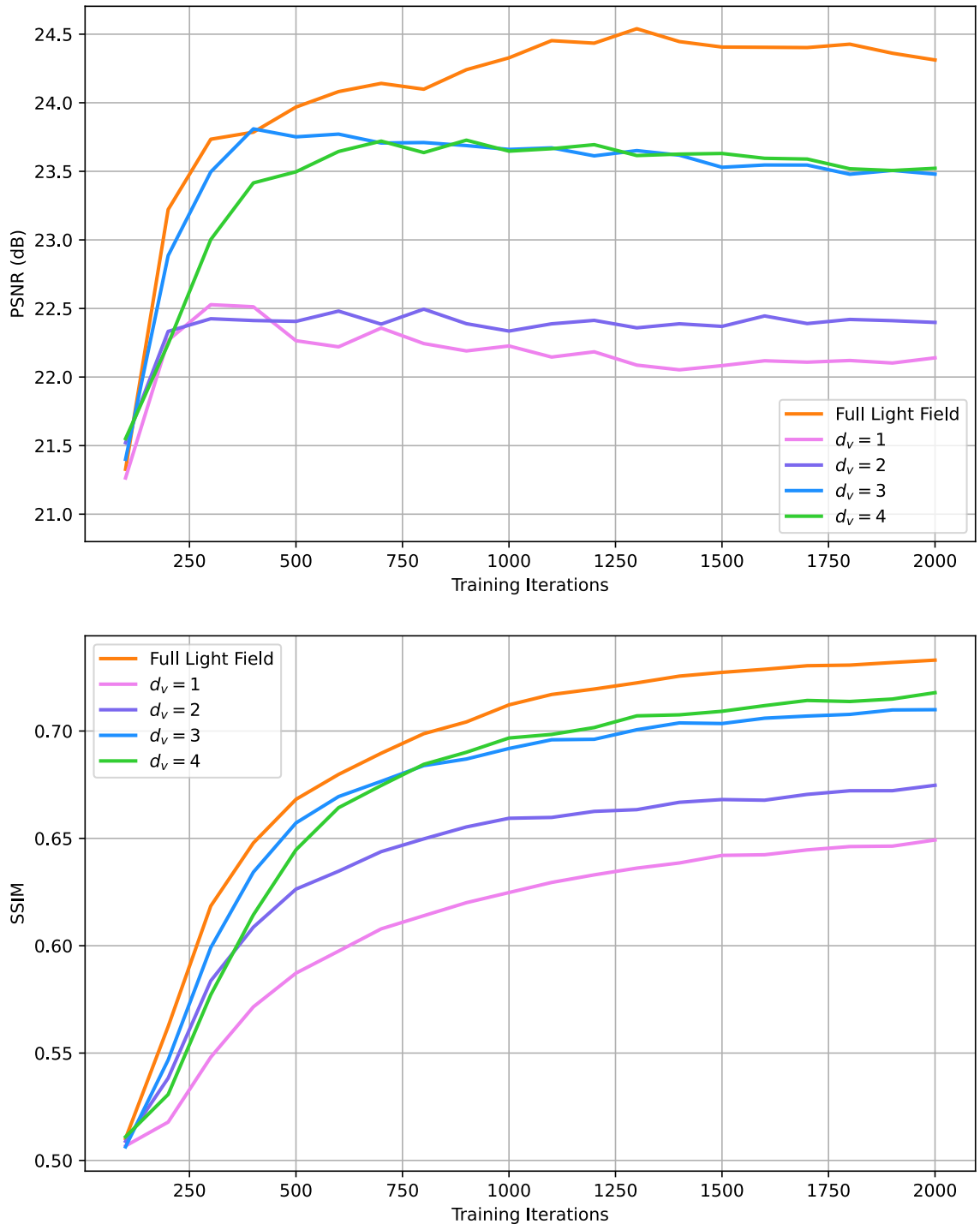

FIGURE 5.13. **Training convergence for different view-based sampling methods using** $s = 4/17$**.** From both PSNR (top) and SSIM (bottom) plots, we see that performance increases as the view distance d_v increases. For $d_v = 4$, performance on SSIM is close to full light field performance, despite being trained on less than a quarter of the data.

are not obvious, the right bird is reconstructed differently with varying view distance. The right birds are reconstructed similarly for $d_v = 3$ and $d_v = 4$, though these reconstructions are considerably better than the reconstructions for $d_v = 1$ and $d_v = 2$. Additionally, the ghosting artifact between the two birds appears less prominent with increasing view distance.

5.6 Convergence

We also examine the training convergence, shown in Figure 5.13. We see that performance on both PSNR and SSIM increases with increasing view distance. With the larger view distance values of $d_v = 3$ and $d_v = 4$, SSIM performance is close to full light field performance, despite training on less than a quarter of the data. Hence, sampling subviews with larger view distances retains more useful information about the scene for view synthesis tasks.

TABLE 5.3. **Reconstruction performance of view-based sampling for varying view distances.** These results show that converged performance increases with greater view distances.

Sampling Method	Number of Rays	PSNR	SSIM
Full Light Field	1,671,168	24.31	0.7330
$d_v = 1$	393,216	22.13	0.6492
$d_v = 2$	393,216	22.40	0.0.6747
$d_v = 3$	393,216	23.48	0.7100
$d_v = 4$	393,216	**23.52**	**0.7179**

CHAPTER 6

Discussion

In this chapter, we derive the key insights and implications of our work, and discuss how these address the knowledge gaps and limitations present in the current literature.

6.1 View Synthesis from Light Field Imaging

Our work developed a novel light field image view synthesis pipeline, which leveraged and adapted NeRF to reconstruct scenes from a set of light field source views. The application of NeRF for light field image inputs is a novel area of research which has not been addressed by existing literature.

With our pipeline, we also addressed the limitations of existing view synthesis techniques in reconstructing scenes when only few or limited views are available. In this challenging scenario, even state-of-the-art methods such as NeRF [**3**] fail to converge to reasonable results when applied on conventional images. As the amount of measured data of a scene is reduced, performance of conventional imaging techniques rapidly degrade, as shown by the few-shot conventional imaging reconstruction in Figure 5.8. This is not an inherent flaw of the view synthesis method, but instead a result of having insufficient measured information of the scene.

Existing solutions in the research literature addressed this information deficiency by leveraging learnt priors [**19, 20**] over a set of previous scenes. However, as discussed in our literature review, these methods introduce various drawbacks. In this work, we instead addressed this information deficiency by targeting the limitations of the conventional method of data capture

(using conventional cameras), as we recognise that this presents a fundamental limiting factor for view synthesis performance, particularly in the few-shot scenario.

Cameras image a scene by measuring the intensity values of light rays intersecting pixels on a camera sensor, where each ray is first focused through a camera lens. Conventional cameras solely image in spatial resolution, where every ray measurement corresponds to a different point in the scene. Our results show that this type of conventional data capture is far from optimal for view synthesis tasks. We demonstrated that a light field camera, which images in both spatial and angular resolution, is much more adept towards performing view synthesis and reconstruction, even when using the same number of ray measurements of a scene. This is particularly evident in the result showing training convergence for the few-shot scenario in Figure 5.10. In this result, the light field pipeline uses the same number of measured light rays of the scene as the conventional imaging pipeline, but achieves drastically higher reconstruction performance on both PSNR and SSIM metrics, as well as much higher observed visual quality (from Figure 5.8). With the same number of ray measurements of the scene, the light field pipeline is able to accurately learn the scene's appearance and geometry, while the conventional pipeline fails to converge, producing a geometrically incoherent reconstruction.

The primary advantages of our pipeline in reconstruction quality were realised in the few-shot scenario, where limited source views were available. However, while in the multi-view scenario reconstruction quality was comparable across both pipelines, the light field pipeline saw a much faster rate of convergence during the training process. For example, in the training convergence shown in Figure 5.6, the view-based sampling light field approach surpassed the converged conventional imaging performance in less than half the number of training iterations.

Hence, a key insight of our work is demonstrating how changing the way a scene is measured from an imaging device can have a significant impact on learning and understanding the scene, and performing view synthesis reconstructions. In this regard, light field cameras measure the scene in a way which facilitates better understanding of scene geometry in comparison to conventional cameras, leading to better reconstruction quality in view synthesis, particularly in challenging contexts such as few-shot reconstruction.

6.2 Ray Sampling

We introduced four different ray sampling methods: uniform sampling, random sampling, view-based sampling, and image gradient sampling.

Uniform and random sampling are more naive approaches, which aim to subsample by sampling an uniform spatial distribution of the entire light field image. Interestingly, random sampling slightly outperforms uniform sampling, demonstrating that the introduction the addition of randomness in subsampling can provide beneficial performance advantages.

View-based sampling exploited the intuition of sampling rays which cover a greater difference in angular resolution, where light field subviews further from the centre subview were sampled more densely. This achieved higher performance in comparison to all other proposed sampling approaches. After further evaluating sampling of various subviews, our results showed that as subviews with greater baselines were selected, reconstruction quality increased (Figure 5.13).

We implemented image gradient sampling to introduce a semantic sampling method, which samples rays depending on the observed pixel values. Unfortunately, this method did not achieve any performance gain over the more naive sampling methods, and even performed worse in many scenarios on the evaluation metrics. One possible reason for this is that the intuition behind this sampling approach does not align with the actual sampling which would maximise performance on the evaluation metrics. NeRF is able to converge to better solutions when points in a scene are observed multiple times by different captured light rays. It is possible that by focusing ray sampling on areas with finer detail and texture using image gradients, the overall scene structure and geometry is neglected in comparison to a uniformly-distributed sampling, leading to slightly worse convergence in some scenarios. In particular, for the scenarios we evaluate in, good performance on the evaluation metrics can be achieved simply by converging to geometrically-accurate scene representations, without necessarily requiring accurate reconstruction of fine detail and textures. However, this reasoning is purely speculation, and may not be the true reason for this sampling approach performing worse.

Overall, our development of ray sampling methods allow our view synthesis pipeline to be optimised for different trade-offs between reconstruction quality and computational cost. Different operating scenarios and scenes will determine what the optimal sampling rate is to achieve the desired performance.

6.3 Optimal Light Field Camera Design

As our work found various fixed pattern ray sampling methods which perform better in data fidelity than naively using a uniform distribution of rays across light field subview, this has many implications on optimising hardware design for light field cameras. If we find a particular fixed sampling pattern to be superior, then we should directly attempt to implement it at a hardware level, as downsampling rays in software post-capture is a waste of imaging resources if the same rays are being discarded every time.

Existing research literature has demonstrated many applications where using the full number of subviews in light field images is largely redundant. Jiang *et al.* [**22**] and Kalantari *et al.* [**21**] addressed different computer vision tasks using light field imaging, presenting techniques which used fewer subviews of the light field with minimal sacrifice to performance. In this work, we extended upon these insights by evaluating light field subsampling on a per-ray and per-subview basis, focusing on the specific application towards novel view synthesis.

While sampling on a per-ray basis allows for a more fine-grained evaluation of which ray measurements in the light field are inherently more useful, translating an arbitrary ray sampling pattern to an actual hardware design of a camera is largely non-trivial. For example, the linear distance-weighted view-based sampling pattern presented in Figure 4.4 could be realised in hardware by building a sparse light field camera where different subviews use cameras of different spatial resolutions. Specifically, higher resolution cameras could be placed at positions further from the centre subview to achieve a measure rays in a similar pattern. For more complex ray sampling methods, it may be infeasible to translate the ray sampling pattern to a similar implementation in hardware. Fortunately, from the results in our work, we encounter maximum performance when sampling on a per-subview basis rather than a

per-ray basis, so discussions on how more complex ray sampling patterns might be translated to camera hardware designs can be left as a path for future research.

Our results showed that view-based sampling achieved higher data fidelity than uniformly sampling rays at each camera subview. When discretely sampling on a per-subview basis, we qualitatively demonstrated and evaluated the relationship between camera baselines (i.e. distances between subviews) and view synthesis reconstruction quality. Notably, Figure 5.13 in our results shows the increase in reconstruction quality which arises from exploiting subviews which are further apart from each other.

Thus, our recommendation for light field camera design builds upon the insights drawn by the work of Jiang *et al.* [**22**] and Kalantari *et al.* [**21**] in exploiting specific light field subviews. We observed a larger increase in data fidelity from sampling rays originating further from the centre subview. This implies that instead of adding additional cameras between the corner subviews for a sparse light field camera to further increase the angular resolution, it is more useful to instead equivalently increase the spatial resolution of the corner views to maximise data fidelity.

Therefore, from our results, we recommend sparse light field camera design simply using the four corner subviews furthest from the centre, with emphasis on maximising the camera baseline within the bounds of practicality for the operational context. This design has the added benefit of enabling more economical hardware construction in comparison to a sparse light field camera such as the EPIModule used in our experiments. However, it is important to recognise that this recommendation is derived from the evaluation scenarios and scene conditions used in our work. It is possible that this may not translate to all scenarios and different scene conditions, which is a possible direction for future evaluation.

6.4 Memory/Bandwidth Constraints

Our work addresses applications of view synthesis with constrained computational memory or bandwidth through two considerations: data fidelity optimisation, and fixed pattern sampling.

By employing ray sampling in our pipeline, we allow for optimisation of the trade-off between computational memory cost and reconstruction quality. We show that certain sampling methods, such as view-based sampling, achieve high performance on data fidelity, such that downsampling with larger sampling rates still preserves a significant amount of useful information for reconstruction. In particular, this will enable low-bandwidth view synthesis applications, where the amount of data being transmitted is limited. Consider an operational scenario where a sensor is communicating data on a limited bandwidth to an external source which then performs view synthesis or reconstruction of the scene. In this scenario, by using our light field imaging pipeline, significantly less data is required to be transmitted to achieve high performance.

Our development of fixed pattern sampling variations also further facilitate operation in bandwidth-limited scenarios. For scenarios where bandwidth or computational memory is limited, variable pattern sampling methods are not desirable. If the ray sampling pattern changes between views, then then the data being communicated must also transmit the information on the ray sampling used for every captured view. If the ray sampling pattern is predetermined and constant across views, this is not necessary; the ray sampling pattern only needs to be transmitted once. In our results, we showed that fixed pattern sampling variations perform on par with their variable pattern counterparts in reconstruction quality. Therefore, we recommend use of the fixed pattern sampling variations for all view synthesis scenarios to optimise memory performance with no expense to reconstruction quality.

6.5 Practicality

Ultimately, the implementation practicality of our work for real-world applications is dependent on accessibility to light field imaging hardware. Having observed the complex multi-camera array imaging setups shown in this work to capture light field images, one might question the feasibility and practicality of implementing these techniques due to hardware accessibility and costs.

Therefore, it is worth emphasising that light field cameras are not necessarily more expensive than conventional cameras. A conventional camera can be readily repurposed to capture light field images through the addition of a lenslet array [**12**], which can be produced at low cost. Building a light field camera this way directly trades off the higher spatial resolution of a conventional camera for the higher angular resolution of a light field camera. In our experimental setup, since we evaluated on a multi-camera array light field camera, this direct trade-off was not present in our hardware, since multi-camera array constructions introduce angular resolution by simply adding more cameras, with no expense to the spatial resolution of each subview. However, we evaluated this trade-off fairly by spatially downsampling each of the light field subviews such that both cameras measured the same amount of data, and consequently demonstrated that light field images perform better on view synthesis data fidelity: data from a light field view is more useful than than the equivalent amount of data from a conventional camera view.

Assuming that light field cameras can be constructed with minimal cost overhead to a conventional camera with the equivalent imaging resolution, our work then demonstrates that the performance of conventional camera view synthesis can be matched and in many cases surpassed by using a lower-resolution, lower-cost light field camera in combination with our view synthesis pipeline.

CHAPTER 7

Conclusion

In this work, we developed a light field image view synthesis pipeline, which leverages and adapts NeRF to learn a scene representation and synthesise novel views of a scene. This addressed the underperformance of existing view synthesis methods which relied on conventional camera imaging: we demonstrated that our pipeline achieves better reconstruction quality and faster convergence compared to the equivalent state-of-the-art conventional image NeRF pipeline. In particular, we showed that in few-shot view synthesis scenarios, our pipeline is able to produce accurate reconstructions of scene appearance and geometry where conventional NeRF reconstructions fail and produce incoherent results.

Through our development of different light field ray sampling techniques to optimise data fidelity, we facilitate low-bandwidth and memory-constrained operational scenarios for view synthesis, and provide insight into which rays within are light field are inherently more useful. This provides implications on how light field cameras can be better optimised in their hardware design, and consequently we are able to propose light field camera designs which may be more optimal for view synthesis applications.

Ultimately, these contributions enable the realisation of higher-performance, lower-cost imaging setups for real-world view synthesis applications, encompassing domains such as computer graphics, cinematography, virtual reality and robotics. In robotics, robots will be able to better understand their environments with reduced on-board sensing capacity, and in cinematography and computer graphics, special effects which require view synthesis will be achievable using lower-cost setups with fewer cameras.

7.1 Future Work

In this work, we discussed implications on optimising light field camera designs, based on our results from experimenting with different fixed ray sampling patterns. Our results suggested that a four-view sparse light field camera with good view placement would improve performance in data fidelity over a traditional multi-camera setup with equivalent overall imaging resolution. The continuation of this work would involve constructing this proposed design in hardware, and evaluating its performance in comparison to other light field camera designs to validate the insights we presented.

Another interesting direction for future research which builds upon our work is the further development of more optimal ray sampling approaches. In particular, this work leaves significant space for improvement in the development of better semantic sampling methods. While we developed and implemented a semantic sampling method based on image gradients, this failed to outperform more naive methods such as random sampling. In this work, we derived many ray sampling approaches off intuition. Future work could investigate a more methodical optimisation approach, such as learning ray sampling patterns using supervised machine learning techniques.

References

[1] S. T. Digumarti, J. Daniel, A. Ravendran, R. Griffiths, and D. G. Dansereau, "Unsupervised learning of depth estimation and visual odometry for sparse light field cameras," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 278–285, IEEE, 2021.

[2] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, pp. 102:1–102:15, July 2022.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[4] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[5] J. Tremblay, M. Meshry, A. Evans, J. Kautz, A. Keller, S. Khamis, C. Loop, N. Morrical, K. Nagano, T. Takikawa, *et al.*, "RTMV: A ray-traced multi-view synthetic dataset for novel view synthesis," *arXiv preprint arXiv:2205.07058*, 2022.

[6] N. Zeller, F. Quint, and U. Stilla, "From the calibration of a light-field camera to direct plenoptic odometry," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 7, pp. 1004–1019, 2017.

[7] D. G. Dansereau, I. Mahon, O. Pizarro, and S. B. Williams, "Plenoptic flow: Closed-form visual odometry for light field cameras," in *2011 IEEE/RSJ international conference on intelligent robots and systems*, pp. 4455–4462, IEEE, 2011.

[8] D. Tsai, D. G. Dansereau, T. Peynot, and P. Corke, "Image-based visual servoing with light field cameras," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 912–919, 2017.

[9] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light field features in scale and depth," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2019.

[10] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 31–42, 1996.

[11] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *ACM SIGGRAPH 2005 Papers*, pp. 765–776, 2005.

[12] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford University, 2005.

[13] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[14] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 43–54, 1996.

[15] S. Liu, T. Li, W. Chen, and H. Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7708–7717, 2019.

[16] W. Chen, H. Ling, J. Gao, E. Smith, J. Lehtinen, A. Jacobson, and S. Fidler, "Learning to predict 3d objects with an interpolation-based differentiable renderer," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, "Deepview: View synthesis with learned gradient descent," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2367–2376, 2019.

[18] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[19] V. Sitzmann, S. Rezchikov, W. T. Freeman, J. B. Tenenbaum, and F. Durand, "Light field networks: Neural scene representations with single-evaluation rendering," in *Proc. NeurIPS*, 2021.

[20] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelNeRF: Neural radiance fields from one or few images," in *CVPR*, 2021.

[21] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–10, 2016.

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

[23] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.