UNIVERSIDADE
BEIRA INTERIOR

Light Field Compression and View Synthesis
An Objective Quality Analysis

Daniela Ferreira Saraiva

Dissertação para obtenção do Grau de Mestre em
Engenharia Eletrotécnica e de Computadores
(2º ciclo de estudos)

Orientador: Prof. Doutor António Manuel Gonçalves Pinheiro

Outubro de 2024.

Declaração de Integridade

Eu, Daniela Ferreira Saraiva, que abaixo assino, estudante com o número de inscrição M12941 de/o Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o **Código de Integridades da Universidade da Beira Interior**.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 11 /10 /2024

Assinado por: **DANIELA FERREIRA SARAIVA**
Num. de Identificação: 15548361
Data: 2024.10.11 19:41:13+01'00'

Resumo

A tecnologia *light field* representa uma tecnologia de imagem poderosa que capta a intensidade e direção dos raios de luz num cenário, permitindo a reconstrução de informação 3D e a realização de tarefas exclusivas, como a refocagem de imagens após a captura. No entanto, a grande quantidade de dados gerada por esta tecnologia apresenta desafios significativos em termos de armazenamento e transmissão, sendo necessário o desenvolvimento de esquemas de compressão eficientes para lidar com estas exigências. As técnicas de compressão que incorporam síntese de vistas durante diferentes fases do processo de compressão surgiram como uma solução promissora, reduzindo a quantidade de dados que precisam ser transmitidos ou armazenados ao permitir a reconstrução de novas vistas a partir de um conjunto limitado de dados capturados. Nesta tese será avaliada uma versão deste método.

O principal objetivo desta tese é estudar o potencial da utilização de síntese de vistas para melhorar a compressão de *light fields*, reduzindo a complexidade, e com foco em manter uma qualidade de imagem alta ao reduzir os requisitos de armazenamento e transmissão de dados.

Para além deste objetivo principal, existem alguns objetivos secundários relacionados com as ferramentas utilizadas neste processo. É efetuada uma comparação entre os codecs utilizados para analisar o impacto que a escolha de codec tem na qualidade da reconstrução final quando lhe é adicionado a técnica de síntese de vistas. Ao focar num único método de síntese de vistas de última geração, a sua capacidade para reconstruir imagens de alta qualidade a partir de dados comprimidos vai ser testada. Além disso, são utilizados quatro *light fields*: dois do conjunto de dados da EPFL, que consiste em *light fields* capturados por câmaras lenslet, e dois da base de dados de *light fields* HCI, que consiste em *light fields* criados sinteticamente. Esta diversidade tem como objetivo compreender melhor a variabilidade do desempenho e a capacidade de generalização do processo.

Para alcançar isto, um *light field* esparsamente amostrado é criado a partir do *light field* original ao descartar algumas das vistas. Ambos os *light fields*, completo e amostrado, são comprimidos usando os seguintes codecs/configurações: JPEG Pleno, VVC LowDelay e VVC Random Access. Um método de síntese de vistas *learning-based*, SepConv++, é aplicado às vistas descodificadas do *light field* amostrado, obtendo um *light field* reconstruído com as mesmas vistas do original. Tanto o *light field* que foi totalmente comprimido quanto o *light field* esparsamente amostrado, que passa por compressão e síntese de vistas, são comparados ao *light field* original. Esta comparação é feita utilizando as métricas objetivas PSNR-HVS-M, MS-SSIM e FSIMc

Os resultados obtidos foram apresentados sob a forma de quatro tabelas, onde cada uma correspondente a um *light field* específico. Cada tabela contém seis gráficos que ilustram as métricas objetivas PSNR-HVS-M, MS-SSIM e FSIMc para cada codec/configuração

utilizada. Estes gráficos também incluem informações sobre a síntese de vistas, ao apresentar linhas de referência sobre o seu desempenho em *light fields* não comprimidos. Adicionalmente, o processo de síntese de vistas cria três tipos de vistas: as que faziam parte do *light field* esparsamente amostrado e sofreram compressão, as visualizações de primeira geração criadas durante a primeira fase de síntese de vistas e as vistas de segunda geração geradas numa fase subsequente de síntese de vistas. As métricas relativas a estes diferentes tipos de vistas também são retratadas nestes gráficos, para todos os codecs/-configurações. São apresentados resultados adicionais através da métrica de Bjontegaard, uma comparação visual em que são apresentados diferentes tipos de vistas para taxas de bits selecionadas e uma tabela com os tempos de compressão.

A análise dos diferentes codecs demonstrou que o SepConv++ pode efetivamente gerar *light fields* mais densos a partir de vistas comprimidas sem perda significativa de qualidade. Embora os *light fields* amostrados exijam taxas de bits mais baixas para armazenamento e transmissão, para alcançar níveis de qualidade de imagem comparáveis aos dos *light fields* totalmente comprimidos, é necessário utilizar uma taxa de bits semelhante à empregue nos *light fields* totalmente comprimidos. Os resultados evidenciaram que o VVC (em qualquer configuração) supera o JPEG Pleno na preservação da qualidade, embora este último seja significativamente mais rápido. A abordagem proposta, que utiliza vistas amostradas comprimidas para síntese de vistas, mostrou-se vantajosa ao reduzir significativamente a complexidade computacional, especialmente no caso do codec VVC. No entanto, a premissa inicial de que a síntese de vistas permitiria uma redução da taxa de bits mantendo a mesma qualidade não foi totalmente confirmada pelos resultados.

Palavras-chave

Light Field, Compressão, Sintese de vistas, 3D, Armazenamento, Transmissão

Resumo Alargado

A tecnologia *light field* representa uma tecnologia de imagem poderosa que capta a intensidade e direção dos raios de luz num cenário, permitindo a reconstrução de informação 3D e a realização tarefas exclusivas, como a refocagem de imagens após a captura. Os *light fields* são descritos por uma função plenóptica e existem diversas representações desta função. A mais comum é a representação 4D $L(u, v, s, t) \in R^4$, que descreve os raios de luz através da sua interseção com dois planos em posições arbitrárias, onde (u, v) representa o primeiro plano e (s, t) representa o segundo plano.

Ao contrário das câmaras tradicionais, os dispositivos de captura de *light fields* registam não só a luminosidade e cor em cada ponto do sensor, mas também a direção e o ângulo dos raios de luz, permitindo a reconstrução do percurso exato de cada raio. A captura destes pode ser feita utilizando diversos métodos, com destaque nas camaras lenslet. Estas câmaras, para além da lente convencional e do sensor, possuem um conjunto de microlentes que permitem captar uma *raw image*, que, após processamento, é convertida em imagens *sub-aperture*. Estas imagens consistem numa matriz de imagens 2D, conhecidas como vistas, cada uma captura o cenário numa perspectiva ligeiramente diferente. O número de vistas e a sua resolução depende das características de cada câmara. Nesta tese, dois dos *light fields* utilizados foram capturados com uma câmara deste tipo.

Devido à necessidade de armazenar muito mais informação do que uma simples imagem, a grande quantidade de dados gerada por esta tecnologia apresenta desafios significativos em termos de armazenamento e transmissão, sendo necessário o desenvolvimento de esquemas de compressão eficientes para lidar com estas exigências. A investigação em compressão de *light fields* tem vindo a evoluir, focando-se tanto na adaptação de codecs padrão, como o H.264, HEVC e VVC, como em métodos especializados, incluindo o standard plenóptico da Joint Photographic Experts Group (JPEG), o JPEG Pleno.

Este trabalho propõe a avaliação de um método específico que combina a compressão de *light fields* com síntese de vistas, uma técnica que emerge como promissora ao permitir a reconstrução de novas vistas a partir de um conjunto limitado de dados capturados. Este processo, denominado de síntese de vistas, representa um passo importante para melhorar a eficiência de compressão, uma vez que a redução do número de vistas originais visa permitir poupanças consideráveis de memória.

O principal objetivo desta tese é estudar o potencial da utilização de síntese de vistas para melhorar a compressão de *light fields*, reduzindo a complexidade, e com foco em manter uma qualidade de imagem alta ao reduzir os requisitos de armazenamento e transmissão de dados.

Para além deste objetivo principal, existem alguns objetivos secundários relacionados com as ferramentas utilizadas neste processo. É efetuada uma comparação entre os codecs

utilizados para analisar o impacto que a escolha de codec tem na qualidade da reconstrução final quando lhe é adicionado a técnica de síntese de vistas.

Ao focar num único método de síntese de vistas de última geração, a sua capacidade para reconstruir imagens de alta qualidade a partir de dados comprimidos vai ser testada. Além disso, são utilizados quatro *light fields*: dois do conjunto de dados da EPFL, que consiste em *light fields* capturados por câmaras lenslet, e dois da base de dados de *light fields* HCI, que consiste em *light fields* criados sinteticamente. Esta diversidade tem como objetivo compreender melhor a variabilidade do desempenho e a capacidade de generalização do processo.

Para este trabalho o SepConv++ foi escolhido como método de síntese de vistas, embora tenha sido desenvolvido originalmente para interpolação de fotogramas de vídeo, demonstrou um melhor desempenho relativo a outros métodos estado da arte criados exclusivamente para este propósito. Os *light fields* apresentam um número elevado de vistas, o que torna o seu processamento intensivo em termos computacionais, além de demorado. Para otimizar a eficiência e obter resultados para análise, foi decidido focar apenas nas 5x5 vistas interiores. Dois conjuntos de imagens foram criados para o estudo. O primeiro conjunto consiste nas vistas 5x5 selecionadas anteriormente (*light field* 5x5), enquanto o segundo conjunto, amostrado de forma esparsa, é formado por nove vistas retiradas do conjunto inicial (*light field* 3x3).

Ambos os *light fields*, o 5x5 e o 3x3 amostrado, foram codificados a cinco taxas de bits distintas, utilizando os codecs/configurações: Pleno, VVC Low Delay e VVC Random Access. Após a codificação, as imagens foram descodificadas para obter as respetivas vistas.

Ao *light field* amostrado, que foi comprimido, foi aplicado o método de síntese de vistas, de forma a reconstruir um *light field* com o mesmo número de vistas que o original. O processo de síntese de vistas foi realizado em duas fases, resultando num *light field* reconstruído com três tipos distintos de vistas: as nove vistas originais comprimidas, 12 vistas geradas a partir destas (designadas como vistas de primeira geração), e quatro vistas geradas a partir das vistas de primeira geração (designadas vistas de segunda geração).

Por fim, ambos os *light fields* – o totalmente comprimido e o amostrado, que sofreu compressão e posteriormente síntese de vistas – foram comparados com as vistas originais, utilizando métricas objetivas como PSNR-HVS-M, MS-SSIM e FSIMc.

Os resultados obtidos foram apresentados sob a forma de quatro tabelas, onde cada uma correspondente a um *light field* específico. Cada tabela contém seis gráficos que ilustram as métricas objetivas PSNR-HVS-M, MS-SSIM e FSIMc para cada codec/configuração utilizada: Pleno, VVC Low Delay, e VVC Random Access. Estes gráficos permitem comparar o desempenho entre codecs, e tambem o desempenho entre os *light fields* totalmente comprimidos (5x5) e os *light fields* amostrados esparsamente (3x3), que foram reconstruídos através do processo de síntese de vistas.

Os gráficos também incluem linhas de referência que indicam o desempenho da síntese de vistas aplicada aos *light fields* amostrados sem estes serem comprimidos. Estas linhas de referência são fundamentais para avaliar o impacto da síntese de vistas, avaliando a qualidade do método selecionado (SepConv++) antes de adicionar a componente de compressão.

O processo de síntese de vistas resulta na criação de três tipos distintos de vistas: as vistas do conjunto 3x3 que foram comprimidas, as vistas de primeira geração (criadas na primeira fase do processo de síntese de vistas) e as vistas de segunda geração (geradas numa fase subsequente). As métricas objetivas (PSNR-HVS-M, MS-SSIM e FSIMc) relativas a cada um destes três tipos de vistas também são apresentadas. Estas, fornecem uma visão abrangente sobre como cada fase de geração tem impacto na qualidade das respetivas vistas reconstruidas.

São apresentados resultados adicionais através da métrica de Bjontegaard, uma comparação visual em que são apresentados diferentes tipos de vistas para taxas de bits selecionadas e uma tabela com os tempos de compressão. Todos estes componentes contribuíram para a analise de todo o processo e permitiram retirar as conclusões.

A análise dos diferentes codecs demonstrou que o SepConv++ pode efetivamente gerar *light fields* mais densos a partir de vistas comprimidas sem perda significativa de qualidade. Embora os *light fields* amostrados exijam taxas de bits mais baixas para armazenamento e transmissão, para alcançar níveis de qualidade de imagem comparáveis aos dos *light fields* totalmente comprimidos, é necessário utilizar uma taxa de bits semelhante à empregue nos *light fields* totalmente comprimidos. Os resultados evidenciaram que o VVC (em qualquer configuração) supera o JPEG Pleno na preservação da qualidade, embora este último seja significativamente mais rápido. A abordagem proposta, que utiliza vistas amostradas comprimidas para síntese de vistas, mostrou-se vantajosa ao reduzir significativamente a complexidade computacional, especialmente no caso do codec VVC. No entanto, a premissa inicial de que a síntese de vistas permitiria uma redução da taxa de bits mantendo a mesma qualidade não foi totalmente confirmada pelos resultados.

Abstract

Light field technology represents a powerful imaging technology that captures the intensity and direction of light rays in a scene, allowing for the reconstruction of 3D information and the ability to perform unique tasks like refocusing images after capture. However, the vast amount of data generated by light field imaging poses significant challenges for storage and transmission, making efficient compression schemes crucial. Compression techniques that incorporate view synthesis during different stages of the compression process have emerged as a promising solution, reducing the amount of data that needs to be transmitted or stored by reconstructing or predicting new views from a limited set of captured data. In this thesis a version of this method will be evaluated.

The main goal of this thesis is to study the potential of using view synthesis to improve light field compression while reducing complexity, and focusing on maintaining high image quality while reducing data storage and transmission requirements.

In addition to this main goal, there are some secondary objectives related to the tools used in this process. A comparison between the used codecs is done to analyze how the choise of codec impacts the final reconstruction quality when coupled with the view synthesis technique. By focusing on a single, state-of-the-art view synthesis method, its ability to reconstruct high-quality images from compressed data is tested. Additionally, four light fields are used: two from the EPFL dataset, which consists of light fields captured by lenslet cameras, and two from the HCI Light Field Database, consisting of synthetically created light fields. This diversity aims to better understand the variability in performance and the generalization capability of the process.

To achieve this, a sparsely sampled light field is created from the original light field by "dropping" views. Both light fields, complete and sampled, are compressed using the following codecs/configurations: JPEG Pleno, VVC LowDelay and VVC Random Access. A learning-based view synthesis method, SepConv++, is applied to the decoded views from the sampled light field, obtaining a reconstructed light field with the same views as the original. Both the fully compressed light field and the sparsely sampled light field, which undergoes compression and view synthesis, are compared to the original light field. This comparison is done using the objective metrics PSNR-HVS-M, MS-SSIM and FSIMc.

The results obtained were presented in the format of four tables, one for each light field, with each table containing six plots that illustrate the performance metrics PSNR-HVS-M, MS-SSIM, and FSIMc for every codec/configuration used. This plots also include information regarding the view synthesis, by presenting reference lines regarding its perfomance on non compressed light fields. Additionally, the view synthesis process creates three types of views, the views that were part of the sparsely sampled light field and underwent compression, the first-generation views created during the first view synthesis

stage, and the second-generation views generated in a subsequent stage of view synthesis. The metrics regarding these different view types, for all codecs/configurations are also depicted in these plots. Additional results are presented through Bjontegaard metrics, a visual comparison where different view types are presented for selected bitrates, and a table containing the compression times.

The analysis across the different codecs, demonstrated that SepConv++ can effectively generate denser light fields from compressed views without significant quality loss. Although sparse light fields require lower bitrates for storage and transmission, achieving comparable image quality levels after view synthesis requires a bitrate similar to the ones used for the fully compressed light fields. The results highlighted that VVC (in either configuration) outperformed JPEG Pleno in quality retention, although the latter is significantly faster. The proposed approach, which utilizes compressed views for synthesis, proved advantageous by significantly reducing computational complexity and resource demands, particularly evident in the VVC codec. However, the initial premise that view synthesis would allow for a reduced bitrate while maintaining the same quality was not fully supported by the results.

Keywords

Light Field, Compression, View Synthesis, 3D, Storage, Transmission

Contents

List of Figures

List of Tables

1D	One-Dimensional
2D	Two-Dimensionsal
3D	Three-Dimensional
4D	Four-Dimensional
4D-PM	4D-Prediction Mode
4D-TM	4D-Transform Mode
5D	Five-Dimensional
AFR	Adaptative Feature Remixing
AR	Augmented Reality
BD	Bjontegaard
BPP	Bits Per Pixel
CNN	Convolutional Neural network
CSF	Constrast Sensivity Function
DCT	Discrete Cosine Transform
EPI	Epipolar Plane Image
FSIM	Feature Similarity Index
FSIMc	Feature Similarity Index for Color Images
GM	Gradient Magnitude
HDCA	Heidelberg Digital Camera Array
HCI	Heidelberg Collaboratory for Image Processing
H.264	Advanced Video Coding
HEVC	High Efficiency Video Coding
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
JPEG	Joint Photographic Experts Group
JVET	Joint Video Exploration Team
LF	Light Field
MLA	Micro-Lens Array
MPEG	Moving Picture Experts Group
MS-SSIM	Multi-Scale Structural Similarity Index
PC	Phase Congruency
PPM	Portable Pixmap Format
PNG	Portable Network Graphics
PSNR	Peak Signal-to-Noise Ratio
RGB	Red, Green, Blue
RD	Rate-Distortion
SAI	Sub-Aperture Image
SSIM	Structural Similarity Index
SR	Super-Resolution

HVS	Human Visual System
VCEG	Video Coding Experts Group
VVC	Versatile Video Coding
VR	Virtual Reality

Chapter 1

Introduction

Light field (LF) technology represents a powerful imaging technology that captures the intensity and direction of light rays in a scene, allowing for the reconstruction of 3D information and the ability to perform unique tasks like refocusing images after capture. However, the vast amount of data generated by light field imaging poses significant challenges for storage and transmission, making efficient compression schemes crucial.

The complexity of light field data has driven extensive research into compression algorithms over recent years, ranging from adaptations of standard video codecs like H.264, HEVC, and VVC to specialized methods tailored for light field data, including a plenoptic coding standard developed by the Joint Photographic Experts Group (JPEG).

While most studies on light field quality-coding focus primarily on visualization, they often overlook the potential benefits of using techniques like view synthesis. Learning-based compression techniques that incorporate view synthesis during different stages of the compression process have emerged as promising, reducing the data that needs to be transmitted or stored by reconstructing or predicting new views from a limited set of captured data, although this topic remains unexplored. These concepts will be explored in this work.

According with some authors, by "dropping views" in the initial stage, compression efficiency can be significantly enhanced, as fewer original views need to be stored or transmitted, resulting in substantial memory savings. View synthesis is then applied to reconstruct the same number of views as the original, ensuring that no information is lost. However, current codecs are very efficient in representing the redundancy between different views, and the improvement in compression efficiency needs to be further researched.

To better illustrate the process, a visual representation of the workflow has been created in Figure 1.1. Compression is applied to both complete light fields (a) and sparsely sampled light fields (b) using the JPEG Pleno, VVC Low Delay, and VVC Random Access codecs/-configurations. View generation is then performed on the decoded views from the sampled light fields. Finally, the output sets are compared to the original light field using objective metrics.

This work specifically evaluates the effectiveness of a single-view synthesis technique applied on the decoder side for light field compression. Most works in quality-compression, focus solely on visualization quality, not considering other light field operations like view synthesis, refocusing or super-resolution. The ability to synthesize opens the door to more

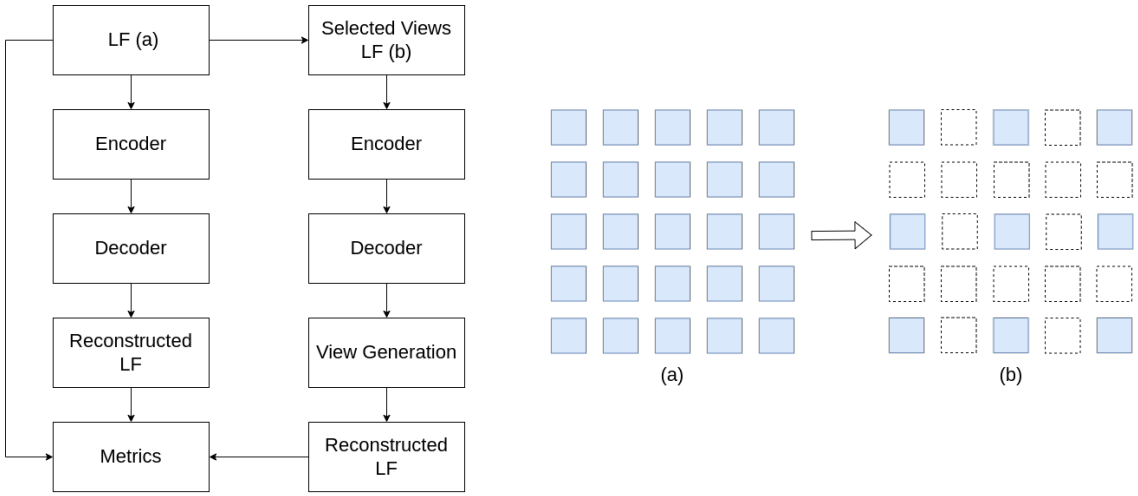


Figure 1.1: Workflow diagram. (a) original light field; (b) sampled light field.

efficient light field coding, allowing for significant memory savings. By reconstructing missing views from a subset of sampled views, it is possible to reduce the data required for transmission or storage, thus optimizing the compression process. The study tests the robustness of this method across four diverse datasets, employing two different codecs (one of these with two different configurations) to provide a comprehensive analysis of how the chosen view synthesis technique interacts with various compression schemes. By focusing on view synthesis, this approach not only enhances visualization but also opens new avenues for efficient light field data coding.

1.1 Scope of this work

The primary goal of this research is to assess the performance of learning-based view synthesis in enhancing light field compression efficiency while maintaining high-quality image reconstruction. The study is structured to achieve the following objectives:

1. **Codec Comparison**: Utilizing two different codecs, one of them with two distinct configurations, to compress light field data, analyzing their impact on the final reconstruction quality when coupled with the view synthesis technique.

2. **View Synthesis Technique**: Focusing on a single, state-of-the-art view synthesis method implemented at the decoder side, examining its ability to reconstruct high-quality images from compressed data.

3. **Performance Metrics**: Evaluating the results using a variety of metrics, including objective measures (such as a modified peak signal-to-noise ratio (PSNR-HVS-M), multi-structural similarity index (MS-SSIM), and feature similarity index (FSIMc).

4. **Dataset Evaluation**: Testing the compression and view synthesis process across

two diverse datasets to understand the variability in performance and generalization capability.

1.2 Thesis structure

This thesis is structured into five chapters, outlined as follows: Chapter 1 presents a brief overview of light field technology, emphasizing one of its challenges and the motivations for this research, along with the objectives and scope of the thesis. Chapter 2 offers a more detailed explanation of light fields, including how they are captured, their unique applications, and a review of the state of the art, with examples of current techniques and uses. Chapter 3 describes the tools utilized in this work and outlines the methodology employed to achieve the results. Chapter 4 presents the results obtained from the experiments along with the discussion and analysis of the results. Chapter 5 consists of conclusions and recommendations for future work and improvements. In Appendix A, additional light field data is presented regarding the definition of parameters in the methodology section.

Chapter 2

State-of-the-Art

2.1 Light fields

The first time the concept of light field was proposed, was in 1936 by Gershun in the publication of a monograph "The Light Field" [16]. The light field was then defined by the totality of rays or radiance in three-dimensional space through any position and in any direction.

$$L : g \rightarrow c \quad (2.1)$$

where:

- L represents the Light field, a function that maps the geometry of light rays to their respective color intensities.
- g refers to the geometry mapping of the light ray, describing its position and direction in space.
- c is a vector describing the intensity of each light component. For example, in the RGB color model, c corresponds to the intensity of red, green, and blue in the light ray.

When talking of light fields, it is essential to introduce the concept of plenoptic function. All 3D representation formats can be obtained through this function, since it describes all visual information. So far, there have been different representations worth mentioning.

Gershun himself defined a five-dimensional plenoptic function $L(x, y, z, \theta, \phi) \in R^5$, where each ray is described by three coordinates (x, y, z) and two angles (θ, ϕ).

In contrast to the previous 5D model, Levoy and Hanrahan (responsible for the introduction of light fields into the computer graphics field) [17] proposed a four-dimensional (4D) representation $L(u, v, s, t) \in R^4$, suggesting that the light field consists of oriented lines in free space. This representation effectively reduces the redundancy in the dataset and simplifies the plenoptic function reconstruction. The 4D representation $L(u, v, s, t)$ describes lines by their intersections with two planes at arbitrary positions, where (u, v) denotes the first plane and (s, t) denotes the second plane (this representations can be seen in Figure 2.1).).

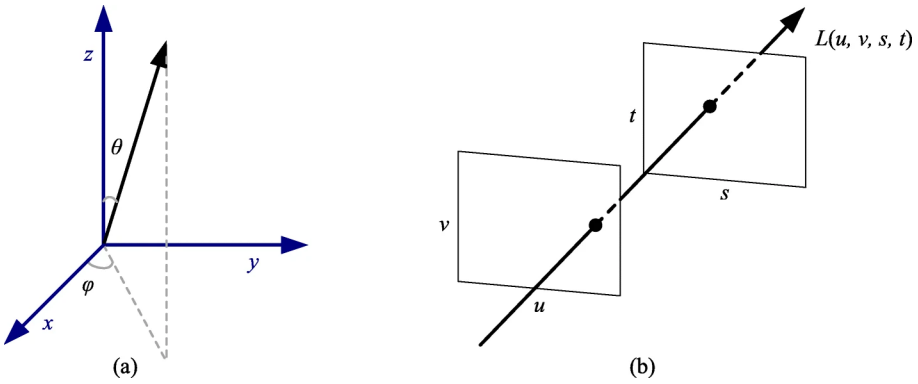


Figure 2.1: 5D and 4D light field representations. a $L(x, y, z, \theta, \phi) \in R^5$, where (x, y, z) represents the coordinates, and (θ, φ) represents the angles between the light ray and the planes; b $L(u, v, s, t) \in R^4$, where (u, v) represents the first plane and (u, v) the second plane[1].

2.2 Light field Acquisition

Unlike traditional cameras that capture a flat, two-dimensional representation of the light rays reaching the lens, light field acquisition devices record not only the brightness and color values at each point on the sensor but also the direction and angle of the incoming light rays. This angular information enables the reconstruction of the exact path each light ray took before reaching the camera sensor. As a result, it becomes possible to compute a three-dimensional model of the scene, providing a more immersive and interactive viewing experience.

2.2.1 Camera Array

Light field acquisition has traditionally been achieved using camera arrays (either by physically moving a single camera to capture static images or by employing a physical $N \times N$ array of cameras), where each camera captures the scene from a different angle and viewpoint, providing a diverse range of perspectives. This method offers high spatial resolution and good image quality, as each camera in the array is equipped with a dedicated $K \times K$ sensor, where K can be a large number. In another words the spatial resolution of the light field data acquired by the camera array is determined by the sensor size of a single camera, and the angular resolution is determined by the number of cameras [18]. The wide baseline created by the physical arrangement of cameras allows for a broad range of object depths to be resolved, enabling detailed 3D reconstructions of the scene [6].

However, the camera array approach comes with several challenges. Synchronizing the shutters of all cameras is essential to avoid temporal discrepancies, and managing varying illumination across different viewpoints can be difficult. Additionally, the processing of vast amounts of data from multiple cameras to generate a coherent light field video is computationally intensive, requiring significant computing power. The wide baseline pro-

vides extensive angular information but also results in a far hyper-focal distance, which, combined with the physical size and complexity of the array, reduces portability and practicality for many applications. Moreover, the spacing between cameras limits the achievable view resolution and the bulkiness of the system poses significant difficulties in terms of synchronization and calibration across multiple cameras ([19], [20]).

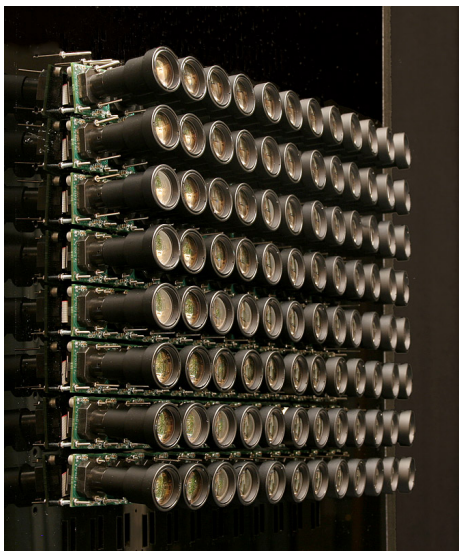


Figure 2.2: The Stanford Multi-Camera Array [2]

2.2.2 Lenslet Cameras

The next significant development in light field imaging that addressed the portability challenges of camera arrays was the invention of the handheld light field camera. This approach uses a single camera with a micro-lens array (MLA) placed between the main lens and the image sensor. The first handheld plenoptic camera, introduced in 2005, utilized this innovative optical design to capture light fields more compactly and conveniently than camera arrays [3].

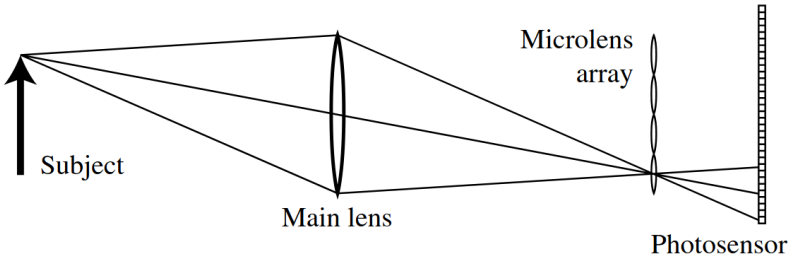


Figure 2.3: The structure of a light field camera,[3]

The light field/plenoptic camera operates through a combination of the main imaging lens, a well-ordered micro-lens array, and a conventional image sensor. As shown in Figure 2.3, the main lens is aligned along the principal optical axis, similar to an ordinary

camera. However, each micro-lens in the array covers a section of the sensor corresponding to a set of pixels, forming what is often referred to as a “macro-pixel” or “lenslet image”. This arrangement allows the camera to capture both the intensity and the directional paths of incoming light rays.

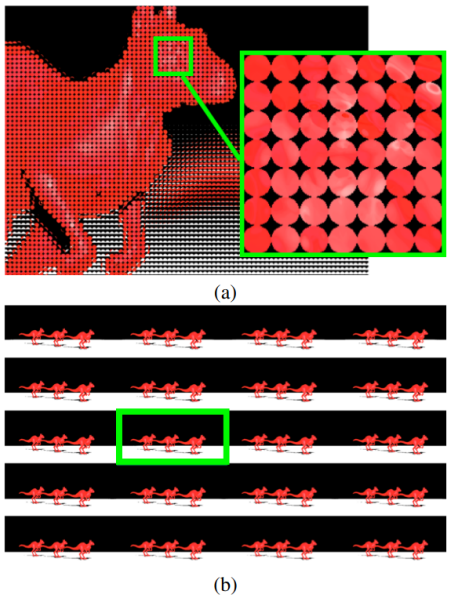


Figure 2.4: (a) Raw light field data. Here each circular image patch (32×32 in this case) corresponds to a single microlens (b) Montage of sub-aperture images (one marked with a green rectangle). Each of these sub-aperture image was generated by sampling the same (u, v) lenslet pixel from every microlens. This format of data essentially is a simple rearrangement of the (a) [4].

Light passing through the main lens is projected onto the sensor after passing through the micro-lens array, forming a unit image for each micro-lens. The resulting raw image data, as shown in Figure 2.4 a), consists of small circular patches corresponding to individual micro-lenses.

By treating each unit image as a macro-pixel, we can extract sub-aperture images (SAIs) (Figure 2.4 b)). These are generated by sampling the same pixel positions across all macro-pixels, yielding an array of images that represent different perspectives of the scene. This method captures both angular and spatial information about the photographed object [18]. The hexagonal macro-pixels in the light field raw data are shown in Figure 2.5 a) and the principle of sub-aperture image extraction in which the pixels in different macro pixels are arranged in order are shown in Figure 2.5 b).

This results in the capturing of $N \times N$ views by the corresponding sensor elements under each microlens, and these sub-images can be processed to produce a set of $N \times N$ images in which each represents a unique view with a spatial resolution of $K \, x \, K$ [6], in Figure 2.6 it is shown how the definition of the parameters K and N change per using a Camera Array or a Lenslet Camera.

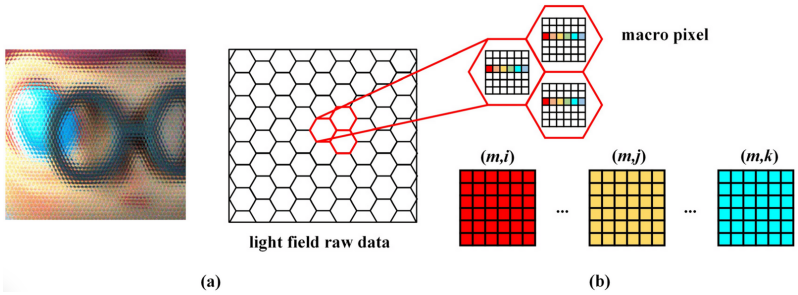


Figure 2.5: Sub-aperture imaging principle. (a) Light field image raw data; (b) Sub-aperture image extraction method [5].

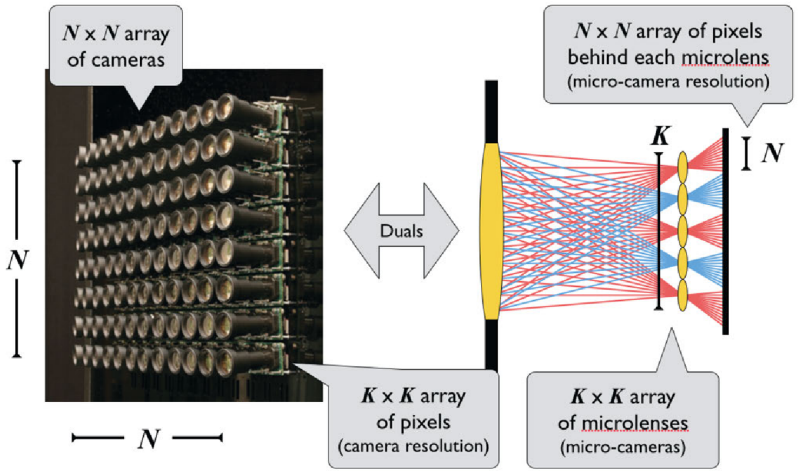


Figure 2.6: Duality between the array-of-cameras approach (left) and the single-camera-with-microlens-array approach (right) [6]

While this setup provides a portable form factor and a high density of views (since N, the number of micro-lenses, can be relatively large for high-resolution image sensors), it comes with a trade-off: the spatial resolution of the final image is limited by the number of micro-lenses, and the very small baseline between left/top and right/bottom views restricts depth imaging to nearby objects.

2.3 Areas of research

The existence of the various unique properties present in light fields leads to multiple areas of research, such as depth estimation, editing techniques, image enhancements, reconstruction, view synthesis, and the light field industry, which includes acquisition and display devices.

2.3.1 Depth Estimation

As mentioned previously, light field data record the spatio-angular details of light rays, making it feasible to extract depth information from these images. Depth cues typically include correspondence cues, defocus cues, binocular disparity, aerial perspective, and motion parallax [1]. When faced with challenges like occlusion—where multiple objects are very close together, obscuring some information—light fields allow for selecting different viewpoints to see past the obstruction, which is particularly advantageous for resolving depth maps with occlusion. Consequently, researchers have focused on various depth estimation techniques, including conventional constraint-based methods [21] , exploring different depth cues and their combinations, simplifying estimation using epipolar image-based methods [22], and applying convolutional neural network (CNN) approaches ([23],[24]) , with special attention to managing occlusion. The key to enhancing other light field-related applications, such as refocusing or rendering, lies in developing more precise and robust depth estimation methods.

2.3.2 Editing

Because light fields are 4D and most tools available are 2D, editing light fields can be difficult. Additionally, local edits must maintain the 4D light field's redundancy, and the 4D light field's implicit depth information makes editing difficult.

Light field editing research areas can be divided into four categories:

- **Refocusing** – Unlike 2D images, light fields allows for refocusing after capture, an example of the application of this technique can be seen in Figure 2.7. Fu et al. implemented the frequency domain digital refocusing of raw images shot with the Lytro light field camera [7].

Figure 2.7: The frequency domain refocusing effects [7]. (a) Raw image, (b) and (c) Images refocused at different focal planes: one focused on the background and the other on the foreground.

- **Removing Occlusions** – Consists in the removal of occlusions. Wang et al. proposed handling the LF de-occlusion (LF-DeOcc) problem using a deep encoder-decoder network "DeOccNet", the first deep learning based method for LF-DeOcc

[8]. They evaluated their proposal by comparing their results to existing methods, and were able to achieve superior performance, as exemplified in Figure 2.8.

Figure 2.8: DeOccNet visual comparison with previous methods, extracted from [8] (a)Occluded (Bike01); (b)Refocusing method [9]; (c)State-of-the art DeOcc method[10]; (d)Proposed DeOccNet[8].

- **Segmentation** – Segmenting the light fields to make the editing experience as smooth as editing a 2D image, simplifying tasks like removing scene objects or changing their color. Wanner et al. [25] developed a globally consistent multi-label assignment framework for light field segmentation. Their approach utilizes appearance and disparity cues to achieve accurate and consistent segmentation across multiple views, leveraging the inherent geometry encoded in light fields to enhance label optimization and improve performance in challenging segmentation scenarios.
- **Interfaces** – Development of tools and software that enable easier editing of all tasks related to light fields. An example of this is LightShop [26], which enables users to interactively manipulate and composite 4D light fields within a unified framework. This system was designed to allow manipulation and rendering of light fields without being constrained by their parametrization or method of acquisition.

2.3.3 Enhancement

There are two main areas of focus when it comes to light-field enhancement, which consists of the optimization of the images quality: deblurring [27] and super-resolution.

During the initial phase of light field super-resolution (SR) research, various 2D image super-resolution methods, including Gaussian models and sparse representations, were applied to light field super-resolution. As deep learning has advanced, approaches for light field image super-resolution that leverage deep learning techniques have become more prevalent and are steadily supplanting traditional techniques [18].

2.3.4 Reconstruction and View Synthesis

When capturing light fields, there is an inherent trade-off when it comes to the spatial and angular resolution that can be obtained, due to hardware limitations. When captured by plenoptic cameras the light field present relatively high angular resolution but a lower

spatial resolution with a more narrow baseline. On the other hand, by using a camera array the result is the opposite, it allows for a larger baseline and higher spatial resolution but lower angular resolution (sparse set of views). Reconstruction/SR methods are presented as a solution for this problem.

To obtain spatial SR from sub-aperture views, there are two approaches that stand out, single-view super-resolution and refinement, and end-to-end residual learning [28].

To overcome the problem of having a sparse set of views (therefore low angular resolution), intermediate views are synthesized between the sub-aperture captured images to obtain dense light fields. The most frequently used methods for view synthesis of light fields include EPI super-resolution, Depth estimation and warping, and multi-plane image generation[28].

Various methods have been developed to perform both spatial and angular super-resolution for input light fields. Earlier approaches to spatio-angular light field reconstruction relied on sparse representation and compressive sensing theories [28]. Recently, deep learning techniques have gained prominence, providing powerful tools for the simultaneous spatial and angular reconstruction of light fields.

For example, Ko et al. proposed a light field super-resolution algorithm based on AFR (adaptive feature remixing). They developed separate SR networks for angular and for spatial super-resolution. These networks use a trainable disparity estimator to extract multi-view features. Then performs feature remixing, these remixed features are then use to reconstruct the images[29]. Also Wu et al. proposed a spatial-angular attention network to perceive non-local correspondences in the light field, and reconstruct high angular resolution light field in an end-to-end manner [30].

2.4 Industry Solutions

The industrial applications of light fields can be divided into two categories, acquisition devices and display devices.

It enables high-resolution imaging, realistic 3D visualization, and immersive experiences without the need for specialized glasses or extensive post-processing. Below, we explore some of the key industrial applications and advancements in light field technology.

These solutions solve problems in two important parts of the light field imaging pipeline: acquisition and display. In the near future, it is expectable the emergence of new portable devices for capturing light fields. In addition, the use of light field displays may extend from fixed screens to display extremely small or extremely large pictures and can further benefit medical microscopy or cinematic displays. Finally, light field technology can contribute to closer-to-truth communication, which should make the “smart life” more attainable.

2.4.0.1 Acquisition

As mentioned previously, the creation and development of handheld light field cameras marked a pivotal moment in the evolution of light field technology, making it more accessible and practical for various applications.

Lytro [31] was a pioneer in this field, introducing the first consumer-grade handheld light field camera. Raytrix [32] followed with high-resolution light field cameras designed for scientific and industrial use, such as microscopy and material inspection, further expanding the practical applications of light field technology.

Wooptix [33] presented a solution to reduce the trade-off between spatial and angular resolution, by using a liquid lens in front of the sensor. This innovative approach allows for rapid changes in focal planes, achieving full sensor resolution in real-time.

Google has also participated in light field acquisition research holding several patents. In 2018, Google introduced a camera design that captures light field images with uneven and incomplete angular sampling [34]. This method improves both spatial resolution and the quality of depth data, which is crucial for applications such as augmented reality (AR), virtual reality (VR), and telepresence.

2.4.0.2 Display

Displaying light fields effectively is another critical aspect of this technology. Several companies are advancing in creating more realistic and immersive display solutions.

Sony's recent advancements include the 3D Spatial Reality Display Technology [35], which tracks the user's eye position to create a glasses-free 3D experience with high resolution. This technology is particularly suited for professional applications like design, medical visualization, and interactive advertising. Another innovation, Atom View ([36], is used in volumetric virtual production through point-cloud rendering. This technology enables the digitization of spaces and objects for virtual sets, which is highly valuable in film production, gaming, and virtual reality experiences.

Looking Glass Factory [37] has developed a light field display capable of providing 45 different viewpoints within a 58° viewing cone. Such technology allows for more dynamic and engaging visual experiences without the need for special glasses, making it suitable for consumer electronics, advertising, and digital art installations.

Light fields have the capability of transforming how we experience digital content and communicate, with Google's Project Starline [38] as a notable example. Designed for real-time communication, Project Starline uses glasses-free light field display technology to create a sense of physical presence during virtual meetings, relying on custom-built hardware and specialized equipment for more natural and engaging interactions in corporate communication and remote collaboration.

2.5 Compression

The complexity of light field data has driven extensive research into compression algorithms over recent years, ranging from adaptations of standard video codecs like H.264, HEVC, and VVC to specialized methods tailored for light field data, including a plenoptic coding standard developed by the Joint Photographic Experts Group (JPEG).

In this thesis both VVC and JPEG Pleno light field codecs will be considered. Because of that they will be further elaborated in the following.

2.5.1 JPEG Pleno

The JPEG Pleno standard aims to establish a comprehensive framework for capturing, representing, and exchanging various plenoptic imaging modalities, such as omnidirectional, depth-enhanced, point cloud, light field, and holographic imaging [39]. Such imaging should be understood as light representations inspired by the plenoptic function, regardless of which model captured or created all or part of the content [39].

In the JPEG Pleno Light Field Coding (Part2) [40], two coding modes are defined, 4D-Prediction mode (4D-PM) and 4D-Transform mode (4D-TM).

While both modes are able to code any light field that is represented as a 2D array of 2D Views (sub-aperture images) [17], the 4D-TM exhibits superior rate-distortion (RD) performance for densely angular-sampled light fields, like lenslet light fields, but struggles with light fields that have broader baselines [12]. On the other hand, the 4D-PM performs better for more sparsely angular sampled light fields at the expense of requiring good quality depth data, while not excelling for denser light fields [12].

This differences in performance are directly related to the functioning of each method. The 4D-PM uses depth-based synthesis, where reference views are encoded using 2D codecs like JPEG 2000 [41], and intermediate views are predicted based on the depth information. As a result, its RD performance is highly dependent on the accuracy of the depth maps. On the other hand, the 4D-TM does not require any depth or geometric data, which makes it more robust in scenarios where such information is not available or is difficult to acquire with sufficient accuracy.

The 4D-Transform Mode (4D-TM) offers a 4D-native coding solution, where redundancies within and across views in the light field are jointly exploited using a multiscale 4D Discrete Cosine Transform (4D-DCT) [12].

The 4D-TM architecture includes five core modules: block partitioning, a transform module, a quantizer, a symbol generator, and an entropy encoder.

It adopts a multiscale 4D-DCT and a hexadeca-tree-oriented bit-plane clustering approach, relying on an adaptive 4D segmentation scheme to partition the light field data into a set

of disjoint four-dimensional (4D) sub-blocks of varied sizes. The set of 4D sub-blocks forms a 4D Partition of the original light field, which is represented by a segmentation tree where each leaf node is associated with a single 4D sub-block. Each 4D sub-block is transformed by a separable 4D-DCT transform and the bit-planes of its coefficients are subsequently encoded using an hexadeca-tree structure [12].

Figure 2.9 illustrates the architecture of the 4D-TM codec, showcasing how the various modules interact to achieve compression.

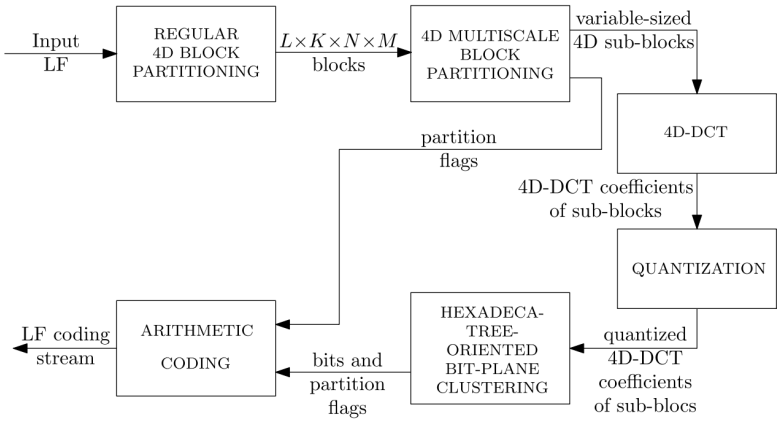


Figure 2.9: JPEG Pleno light field coding 4D-TM architecture according to [11]. Extracted from [12]

.

2.5.2 VVC

The evolution of digital multimedia systems has led to the adoption of hybrid coding architectures, which are particularly effective for compressing light field (LF) images and videos. Among these, Versatile Video Coding (VVC) stands out as a leading codec, developed by the Joint Video Exploration Team (JVET), comprising the ITU-T Video Coding Experts Group (VCEG) and ISO/IEC MPEG [42]. VVC builds on the principles established by its predecessors, such as the High Efficiency Video Coding (HEVC) standard [43], utilizing a combination of prediction and transform coding to efficiently reduce redundancy in visual signals.

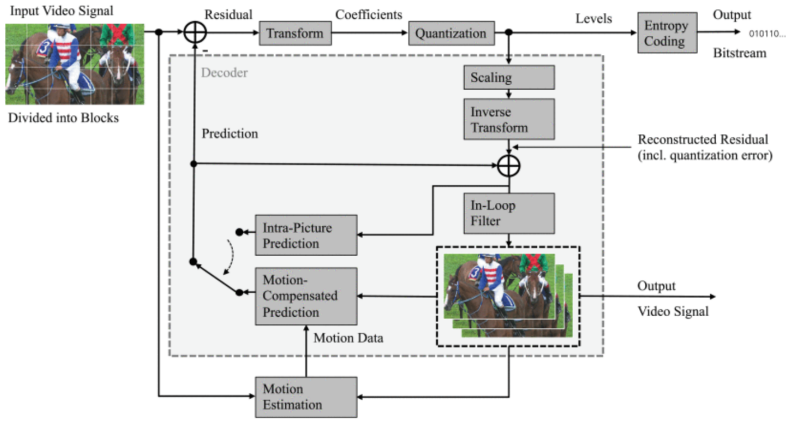


Figure 2.10: Block diagram of a hybrid video encoder, including the modeling of the decoder within the encoder, from [13].

As illustrated in Figure 3.5, a modern hybrid video coder employs several key components, including block partitioning, motion-compensated prediction, and intra-picture prediction [13]. VVC enhances these techniques by introducing flexible partitioning schemes and advanced motion estimation processes, allowing for the efficient encoding of both spatial and temporal redundancies. Furthermore, VVC incorporates innovative transformation and quantization methods, optimizing data representation while minimizing perceptual losses [13].

By leveraging its advanced coding capabilities, VVC presents a promising framework for LF compression, capable of maintaining high fidelity while achieving substantial bitrate reductions.

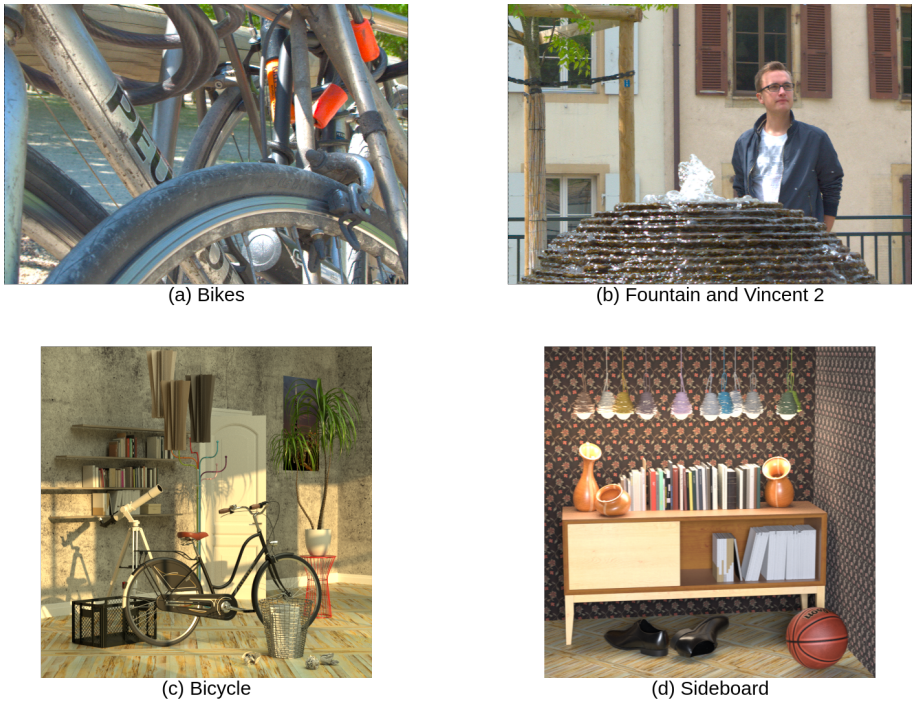
Chapter 3

Methodology

This chapter details the methodology used in the thesis, focusing on the datasets, codecs and reconstruction technique employed, and the integrated workflow. It describes the datasets' characteristics and relevance, the selection and configuration of the codecs, and the step-by-step data processing procedure, including encountered challenges and solutions. This overview aims to clarify the techniques and tools that support the research findings.

3.1 Used Light Fields Datasets

Four light fields were used, to evaluate the performance of using view synthesis for light field compression: Bikes, and Fountain& Vincent2 from the EPFL dataset, and Bicycles and Sideboard from the HCI Light Field Database. The first two were captured by Lenslet Lytro Illum Camera and the other two were synthetically created. The center view for each light field is presented in Figure 3.1. In this study, four datasets were employed to evaluate the performance of light field compression and view synthesis techniques.



(a) Bikes (b) Fountain and Vincent 2

(c) Bicycle (d) Sideboard

Figure 3.1: Center view for every light field used in this work

3.1.1 Lenslet Lytro Illum Camera Datasets

The Lenslet Lytro Illum Camera data set is part of the EPFL dataset [44] [45]. The light field images were captured using a Lytro Illum B01 (10-bit) light field camera.

Both Bikes and Fountain and Vincent 2 were selected from The JPEG Pleno Light Field Datasets according to common test conditions [46]. Bikes belongs to a category named Urban, presenting a high level of spatial information and moderate depth of field. Fountain&Vincent2 is part of the category People, displaying one person and a fountain, which are very close to the camera presenting a high level of spatial complexity.

These light fields sub-aperture images (views) are available in the JPEG Pleno Light Field Datasets as PPM images with RGB color components, non-interlaced. The content is natural and outdoors, and consists of 15×15 views, though only the central 13×13 views are usable to avoid dark views caused by vignetting. Their spatial resolution is 625×434 with a 10-bit depth.

3.1.2 Synthetic HCI HDCA Datasets

These light field images were synthetically created and are made available by the Heidelberg Collaboratory for Image Processing (HCI), with the Bicycle dataset taken directly from 4D Light Field Benchmark website[47].

These light field sub-aperture images are available in the HCI Light Field Database as individual PNG images (views), with RGB color components, non-interlaced. The content consists of synthetic objects, and the dataset provides 9×9 views. Each image has a spatial resolution of 512×512 with an 8-bit depth.

The Sideboard light field from the Synthetic HCI HDCA dataset is also included in the The JPEG Pleno Light Field Datasets according to common test conditions [46], where a modified version of this light field is available. In this version, the views are provided as PPM images with a 10-bit depth, which is particularly useful since the used JPEG Pleno Codec only accepts input images with these characteristics. The version of JPEG Pleno employed here is from its first edition, which had this limitation. Although the current version has resolved this issue, its implementation is not yet publicly available.

3.2 Codecs

3.2.1 JPEG Pleno

As mentioned previously the JPEG Pleno is one of the Codecs employed in this thesis. The software used was “JPEG Pleno Light Field reference software model (4DTM tools)”

[48]. This software provides reference implementations for the standardized technologies within the JPEG Pleno framework for purpose of reference for prospective implementers of the standard and compliance testing.

For this thesis, five different bitrate values were selected for a better analysis of codec performance and interaction with view synthesis. The JPEG Pleno codec and the Bikes light field (a selection of the central 5×5 views set) were used to pick the reference bitrate values.

To vary the bitrate, the parameter λ was used in the encoder input command. This variable represents the Rate-Distortion trade-off: a larger lambda value produces encoded LFs with lower bitrates and thus lower quality, while smaller lambda values result in higher bitrates and better quality.

First, the highest λ value ($\lambda = 20000$), considered the worst acceptable quality based on visual analysis, and the lowest λ value ($\lambda = 200$), where visual analysis indicated no distinguishable difference from the original views, were selected. The corresponding bitrates were used to calculate the three intermediate bitrate values.

3.2.2 Versatile Video Coding (VVC)

Versatile Video Coding (VVC) was finalized in July 2020 as the most recent international video coding standard. It was developed by the Joint Video Experts Team (JVET) of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) to serve an ever-growing need for improved video compression as well as to support a wider variety of today's media content and emerging applications.

The VTM reference software for VVC from Fraunhofer was used. This software package is the reference software for Rec. ITU-T H.266 | ISO/IEC 23090-3 Versatile Video Coding (VVC) [49].

From the VTM software two encoder configurations were used, Random Access and Low Delay. In the encoder configurations files two parameters were set for the required compression. First, for the bitrate control, the QP (quantization parameter) that varies from 0 to 51. Larger QP value produces encoded LFs with lower bitrates and thus lower quality, while smaller QP values result in higher bitrates and better quality. The QP values chosen were the ones that presented the bitrates closest to the reference bitrates.

The second parameter is the InternalBitDepth. It represents the codec operating bit-depth, and it was kept with the value of 10 for all light fields, except for the Bicycle light field, that required a value of 8.

For every light field a configuration file was created with the corresponding characteristics, bellow is the configuration file for the Bikes 5×5 light field to be fully compressed.

```
#======== File I/O ===============
InputFile                : original5x5.yuv
InputBitDepth            : 10         # Input bitdepth
InputChromaFormat        : 444        # Ratio of luminance to chrominance samples
FrameRate                : 30         # Frame Rate per second
FrameSkip                : 0          # Number of frames to be skipped in input
SourceWidth              : 625        # Input  frame width
SourceHeight             : 434        # Input  frame height
FramesToBeEncoded        : 25         # Number of frames to be coded

Level                            : 6.2
```

In Appendix A, the presented tables show the bitrate control parameters defined for every bitrate in each codec/configuration for every light field. For JPEG Pleno, the used parameter is the rate-distortion trade-off (λ), while for VVC, it is the quantization parameter (QP).

3.3 View Synthesis

In light field imaging, when studying the use of view synthesis, the choice of a reconstruction method is vital to achieve high-quality results.

The view synthesis method chosen for this work is SepConv++ [15], an improved version of SepConv [14].

In a comprehensive study Chen et al. [50], considered many types of view synthesis including existing depth image-based rendering, and video frame interpolation methods, that can be directly applied to the LF angular super-resolution problem. Although it was originally created to be applied for video interpolation, SepConv has shown strong performance in light field view generation, outperforming other state-of-the-art methods like Shearlet [51] and LFEPI [19] considering metrics such as PSNR and SSIM.

In 2020 Niklaus et al. [15] described revisiting the 2017 take on SepConv. The paper focuses on improving adaptive separable convolutions for video frame interpolation, achieving near state-of-the-art results by carefully optimizing an older technique. The updated SepConv++ architecture incorporates techniques like residual blocks and kernel normalization, leading to improved interpolation quality. The study highlights the impact of contextual loss and self-ensembling on interpolation quality and suggests future work on high-resolution footage and exploring other applications of adaptive convolutions.

SepConv++ extends the original SepConv neural network architecture, where given input frames I1 and I2, an encoder-decoder network extracts features that are given to four subnetworks that each estimate one of the four 1D kernels for each output pixel in a dense

pixel-wise manner. The estimated pixel-dependent kernels are then convolved with the input frames to produce the interpolated frame Î [14]. An overview of this architecture can be seen in Figure 3.2.

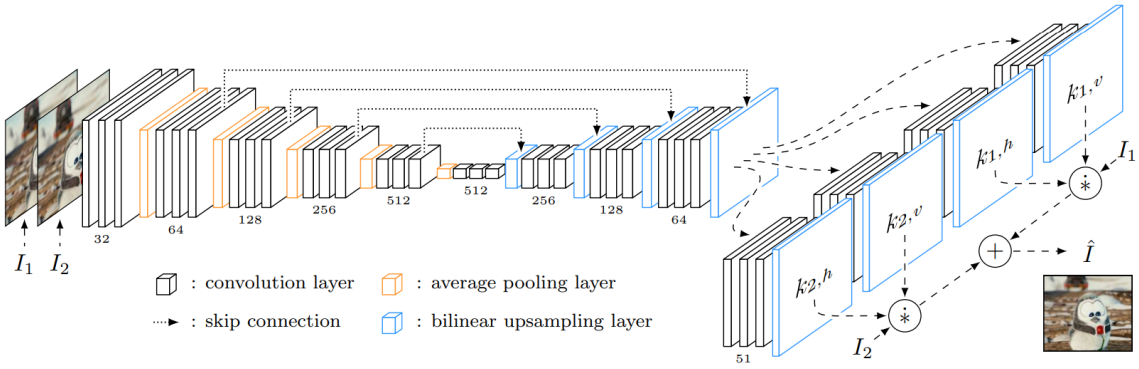


Figure 3.2: Overview of the neural network architecture of SepConv [14]

One of the main enhancements in the updated model (SepConv++) is the inclusion of residual blocks, which take advantage of the significant advancements in deep learning architectures developed after the original release of SepConv. Along with other network improvements, these updates contribute to enhanced interpolation quality.

The kernel normalization strategy was also changed in SepConv++. The updated approach applies adaptive separable convolution to both the input and a mask, and then normalizes by dividing the filtered input by the filtered mask. This modification significantly improves synthesis quality and model convergence. The overview of SepConv++ neural network architecture can be seen in Figure 3.3.

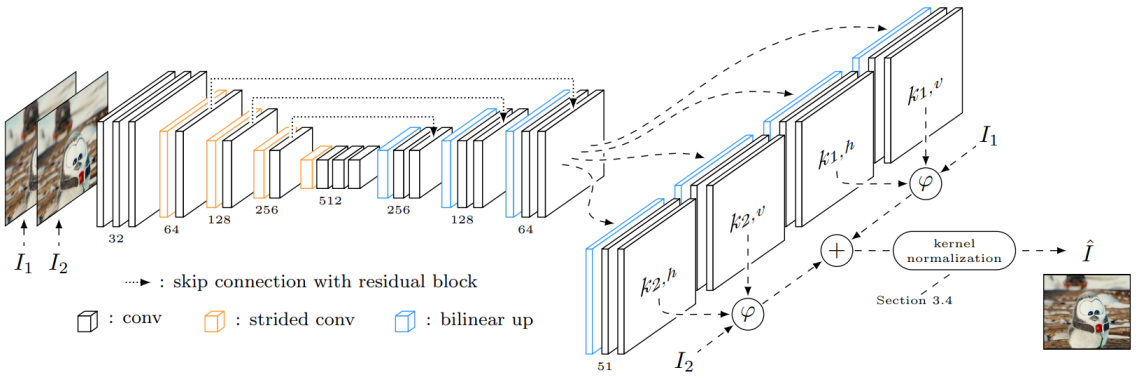


Figure 3.3: An overview of SepConv++ neural network architecture [15]

3.4 Process Workflow

The methodology applied in this work undergoes several modifications depending on the dataset or Codec employed. These differences will be elaborated on later. Nevertheless, the primary methodology is as follows:

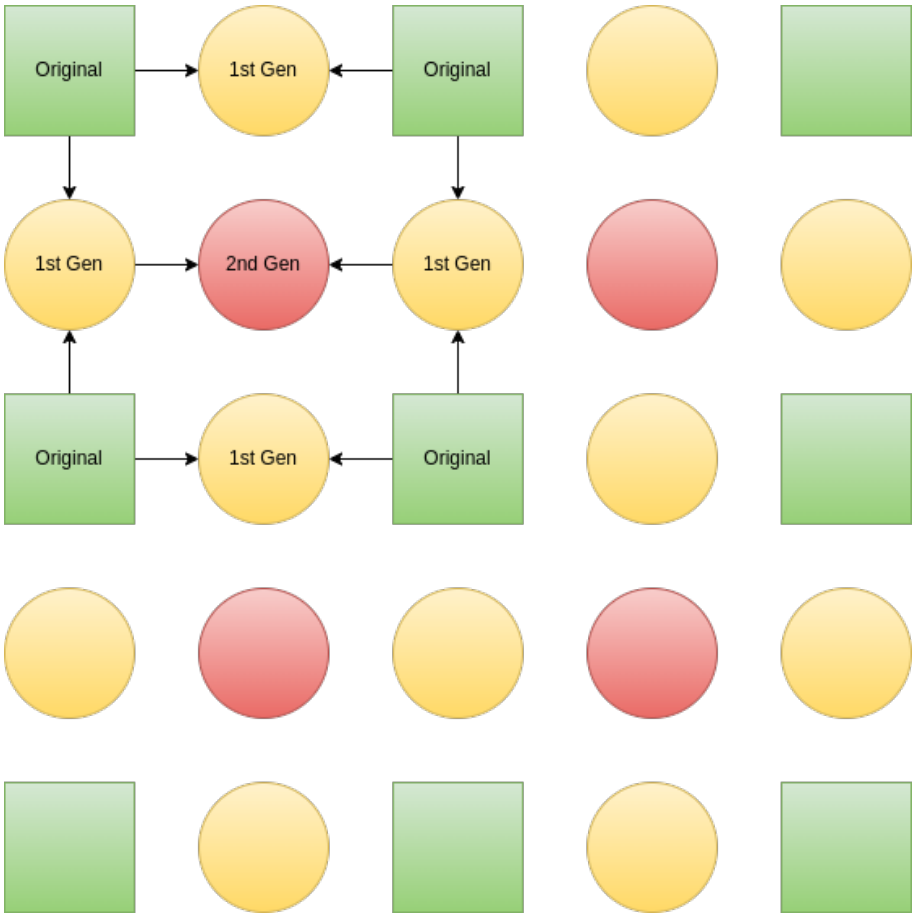


Figure 3.4: View Synthesis process applied to the sparsely sampled 3×3 light field. Legend: **Square** - Original selected views **Yellow Circles** - Reconstructed views in the first view synthesis stage, **Red Circle** - Reconstructed views in the second view synthesis stage.

1. **Selecting the Inner 5x5 Views of a Light field:**
   - Initially, the inner 5x5 views of a light field were selected for further processing. Light fields typically contain a large number of views, which can make the processing computationally intensive and time-consuming, To keep things efficient and get the results for analysis, we focused on using just the inner 5x5 views of the light field for this study .
2. **Creating Two Sets of Images:**
   - One set consists in the selected 5×5 views from the previous step.
   - A new 3×3 sparsely sampled set is made by keeping nine of the views. This selected views are represented by the green squares in 3.4.
3. **Encoding at Different Bit Rates:**
   - Both the 5×5 and 3×3 sets were encoded at five different bit rates using the chosen codecs: Pleno, VVC Low Delay, and VVC Random Access.

- The encoded images were then decoded to obtain the corresponding views.

4. **Preparing the 5x5 Set for Analysis:**
   - The 5×5 fully compressed set, after decoding, was ready to be analyzed for metrics.
5. **Further Processing of the 3x3 Set:**
   - After decoding, view synthesis was applied to the 3×3 views.
   - This resulted in a 5×5 light field with reconstructed views, this process of view synthesis is done in a two stage process that can be seen in 3.4. This reconstructed light field consists of three types of views:
     - 9 original compressed views (green squares)
     - 12 first-generation (yellow circles)
     - 4 second-generation images (red circles)
6. **Comparison with Original Views:**
   - The fully compressed light field and the sparsely sampled that went through view synthesis were then compared with the original views.
   - Metrics such as PSNR-HVS-M, MS-SSIM, and FSIMc were used for comparison.
   - The resulting graphs and tables are discussed in the **Results** section.

As previously mentioned, this process undergoes several adjustments based on the Codec and Dataset being used. These differences will now be explained.

3.4.1 VVC

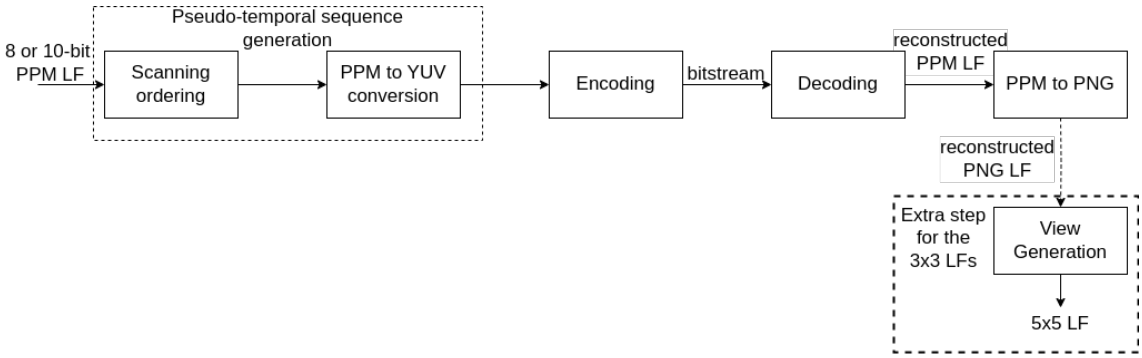


Figure 3.5: Compression process scheme for VVC

The implementaion using VVC is roughly represented in Figure 3.5. The sequence generation was done in a clock wise from inner to outwards manner, this is better described

in Figure 3.6. Its worth mentioning that the conversion from PPM to PNG is necessary to use both the reconstruction technique but also to use the metrics for evaluation. The metrics are used to evaluate the final output ”5×5 LF” which can consist of either 8 or 10-bit PNG light field depending on which dataset is being processed.

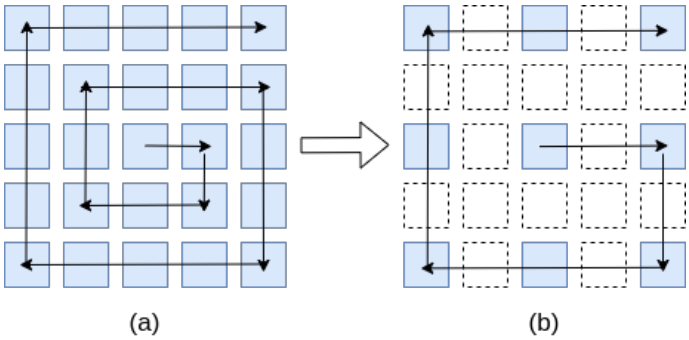


Figure 3.6: Encoding sequence for VVC. (a)- For the original light fields; (b)- For the sampled light fields.

3.4.2 Pleno

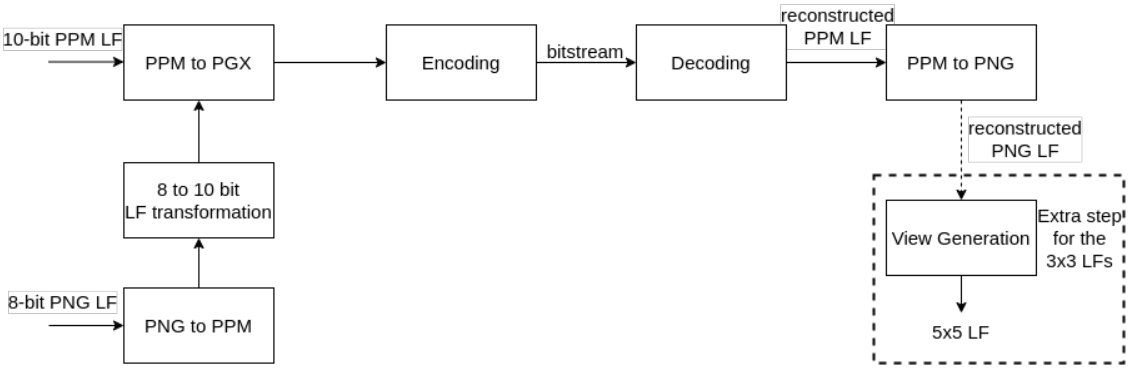


Figure 3.7: Compression process scheme for JPEG Pleno

The implementation using JPEG Pleno is roughly represented in Figure3.7. While implementing this codec it was noticed that the reference software available is not yet prepared to handle light fields with a different bit depth other than 10 bit, that explains the extra steps in the beginning of the scheme to handle the bicycle dataset (the only dataset used in this works that is not part of JPEG Pleno Light Field Datasets, so that presents 8-bit depth) . Once again the conversion from PPM to PNG is necessary to use both the reconstruction technique but also to use the metrics for evaluation. The metrics being applied to the final output “5×5 LF” which consists of a 10 bit depth PNG light field.

3.5 Metrics

Three objective quality metrics were used to compare the result light fields with the orignal ones, PSNR-HVS-M, MS-SSIM and FSIMc. The reference implementation of all this objective quality assessment metrics is available at the "JPEG AI Quality Assessment Framework" repository [52][53].

3.5.1 PSNR-HVS-M

When it comes to image and video processing, PSNR stands out as one of the most widely used objective metrics, primarily due to its simplicity in implementation and its speed of computation. However, while PSNR is an effective metric for comparing different codecs and methods of image compression, its scores do not always correlate well with perceived quality. This limitation arises because PSNR evaluates image quality through a pixel-by-pixel comparison without considering the contextual meaning of the data [54]. This limitation has led to the creation of modified PSNR metrics designed to account for properties of the human visual system.

The PSNR-HVS is an extension of PSNR that incorporates properties of the human visual system (HVS), such as contrast perception ([55],[56]). PSNR-HVS-M further improves upon this by considering visual masking effects ([57],[56]). The matlab implementation of both of this metrics is publicly available [58].

Instead of the standart PSNR, "JPEG AI Quality Assessment Framework" employs the PSNR-HVS-M [57]. It consists of a simple and effective quality model which uses DCT basis functions and is based on the human visual system (HVS). The model operates with 8×8 pixel block of an image and calculates the maximum distortion that is not visible due to the between-coefficient masking [52]. The proposed metric, PSNR-HVS-M, considers the proposed model and the contrast sensitivity function (CSF) [52].

3.5.2 MS-SSIM

The Multi-Scale Structural Similarity (MS-SSIM) [59] is among the most recognized algorithms for assessing image quality, calculating relative quality scores between reference and altered images by comparing details across multiple resolutions. This provides excellent performance for codecs based on machine learning. MS-SSIM offers greater adaptability over single-scale techniques like SSIM, as it accounts for variations in image resolution and viewing conditions. Additionally, the MS-SSIM metric incorporates an image synthesis based method to calibrate the parameters that weight the relative importance between different scales [52]. The smaller the MS-SSIM values the greater the difference between the pixels.

The Multi-Scale Structural Similarity (MS-SSIM) [59] builds on the principles of SSIM by computing the mean, variance, and cross-correlation components at K image scales, where each scale corresponds to low-pass filtering and down-sampling of the original image. The MS-SSIM index is defined as seen in Equation 3.1 [60].

$$\text{MS} - \text{SSIM} = m_K(X,Y)^{\alpha_K} \prod_{k=1}^{K} v_k(X,Y)^{\beta_k} r_k(X,Y)^{\gamma_k} \quad (3.1)$$

Here, $m_k(X,Y)$, $v_k(X,Y)$, and $r_k(X,Y)$ represent the mean, variance, and cross-correlation components computed for patches from scale k. The exponents α_K, $\{\beta_k\}_{k=1}^{K}$, and $\{\gamma_k\}_{k=1}^{K}$ are non-negative and normalize to sum to one across scales (i.e., $\sum_{k=1}^{K} \beta_k = 1$). These vary according to k and adjust the contribution of the components.

3.5.3 FSIMc

The feature similarity (FSIM) metric [61] is based on the computation of two low level features that play complementary roles in the characterization of the image quality and reflects different aspects of the human visual system (HVS): 1) the phase congruency (PC), which is a dimensionless feature that accounts for the importance of the local structure and the image gradient magnitude (GM) feature to account for contrast information [52].

Although FSIM is designed for grayscale images (or the luminance components of color images), the chrominance information can be easily incorporated by means of a simple extension of FSIM, and denominated FSIMc. This is the version used in this thesis [61]. A high metric value express better image quality.

The final FSIM index between two images is given by Equation 3.2, where $S_L(x)$ represents the local similarity map, $PC_m(x)$ is the maximum phase congruency value at each location x, and Ω denotes the entire image spatial domain[61].

$$\text{FSIM} = \frac{\sum_{\mathbf{x}\in\Omega} S_L(\mathbf{x}) \cdot PC_m(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} PC_m(\mathbf{x})} \quad (3.2)$$

The extension for FSIMc involves converting the original RGB images to the YIQ color space, where Y represents luminance and, I and Q represent chrominance. Chrominance similarity, denoted as $S_C(x)$, is calculated using the I and Q components, and the combined chrominance measure is incorporated into FSIMc via Equation 3.3 [61]. In this equation, λ is the parameter used to adjust the importance of the chromatic components in the overall similarity measure.

$$\text{FSIM}_C = \frac{\sum_{\mathbf{x}\in\Omega} S_L(\mathbf{x}) \cdot \left[S_C(\mathbf{x})\right]^{\lambda} \cdot PC_m(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} PC_m(\mathbf{x})} \quad (3.3)$$

Chapter 4

Results

For each light field, six plots were created for analysis, this can be seen in the Tables 4.1, 4.2, 4.3 and 4.4. All of these tables follow the exact same layout, being organized into two columns, each containing a set of three plots. Instead of repeating the legend for each plot, a single legend is placed in the last row of each column, corresponding to the plots above it. This ensures that the legend applies to all three plots in the same column, avoiding unnecessary repetition.

1. **Left column plots**:
    - Three metrics are evaluated PSNR-HVS-M-HVS-M, MS-SSIM and FSIMc (one per plot).
    - A black dashed reference line labeled "ViewGen" can be seen in every plot. This line represents the performance of the view generation method when applied to a sampled light field without compression, serving as a baseline to assess SepConv++ impact.
    - In each plot, there are six curves representing the performance of the codecs employed: Pleno, VVC LowDelay, and VVC Random Access. Each codec/codec configuration is distinguished by color: Pleno is represented in red, LowDelay in blue, and Random Access in green.
    - For each codec/configuration, there are two lines: a continuous line and a dashed line. The continuous line corresponds to the fully compressed 5×5 light field. The dashed line represents the sparsely sampled 3×3 light field, which undergoes compression followed by view synthesis to restore the missing views.
    - This arrangement allows for a clear comparison of the impact of view synthesis on the reconstructed light field quality for each codec.
2. **Right column plots**:
    - Three metrics are evaluated PSNR-HVS-M, MS-SSIM and FSIMc (one per plot).
    - Each plot contains nine curves representing the performance of the different codecs/configurations: Pleno, LowDelay, and Random Access. These are color-coded as follows: Pleno in red, LowDelay in blue, and Random Access in green.
    - For each codec/configuration, there are three lines: a continuous line, a dashed line and a dotted line. These correspond to different view types from the light

field obtained after starting with a 3×3 set and applying view synthesis. The continuous line represents views that were part of the initial view selection and only underwent compression. The dashed line represents the first-generation views created during the initial view synthesis stage. Finally, the dotted line represents the second-generation views generated in a subsequent stage of view synthesis.

- These plots contain two additional reference lines in yellow, which evaluate the view generation process without any compression applied. The yellow dashed line represents the first-generation views created through view synthesis applied to the original 3×3 (no compression applied), and the yellow dotted line corresponds to the second-generation views produced in the subsequent stage.

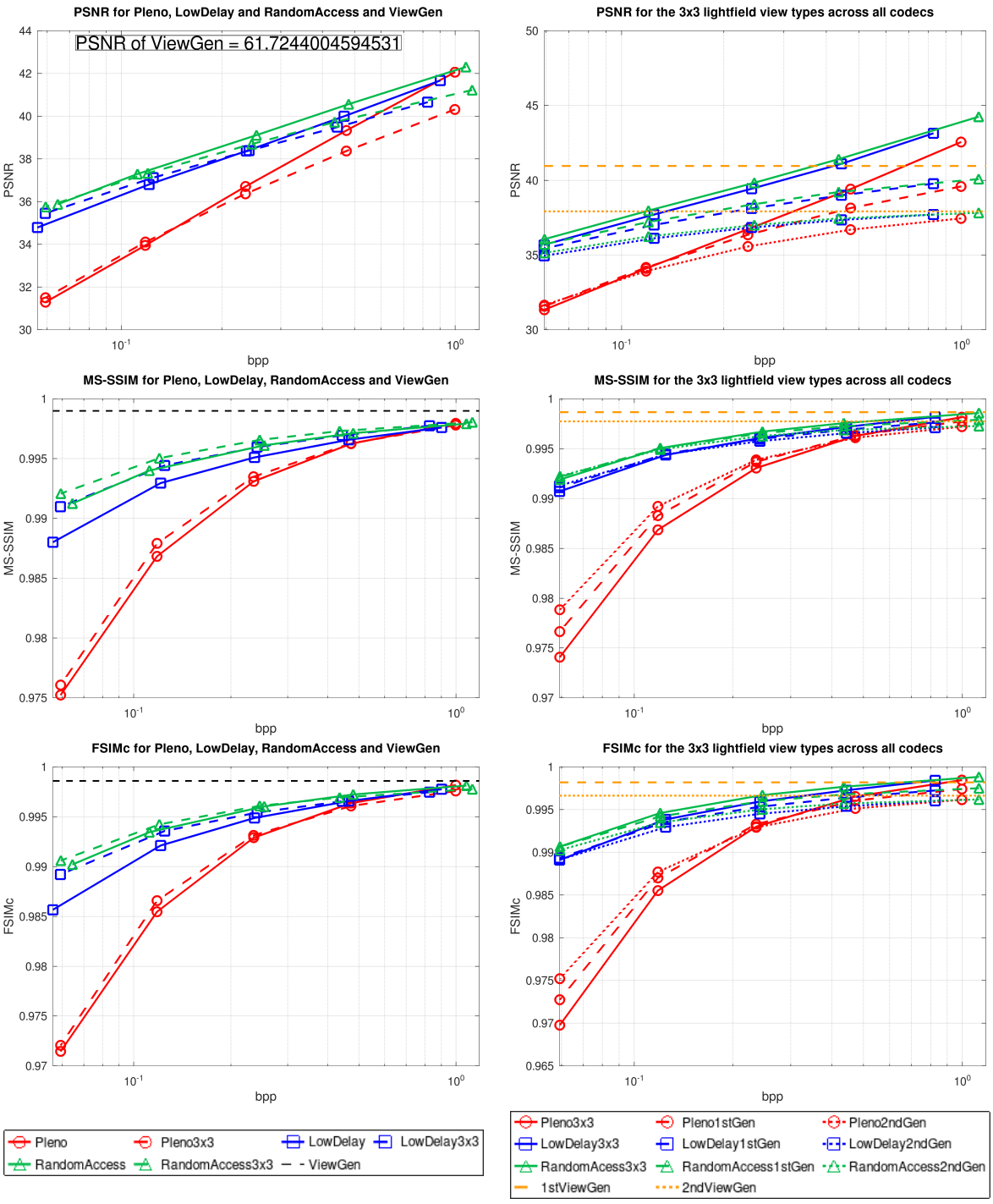


Table 4.1: Plots for the Bikes light field.

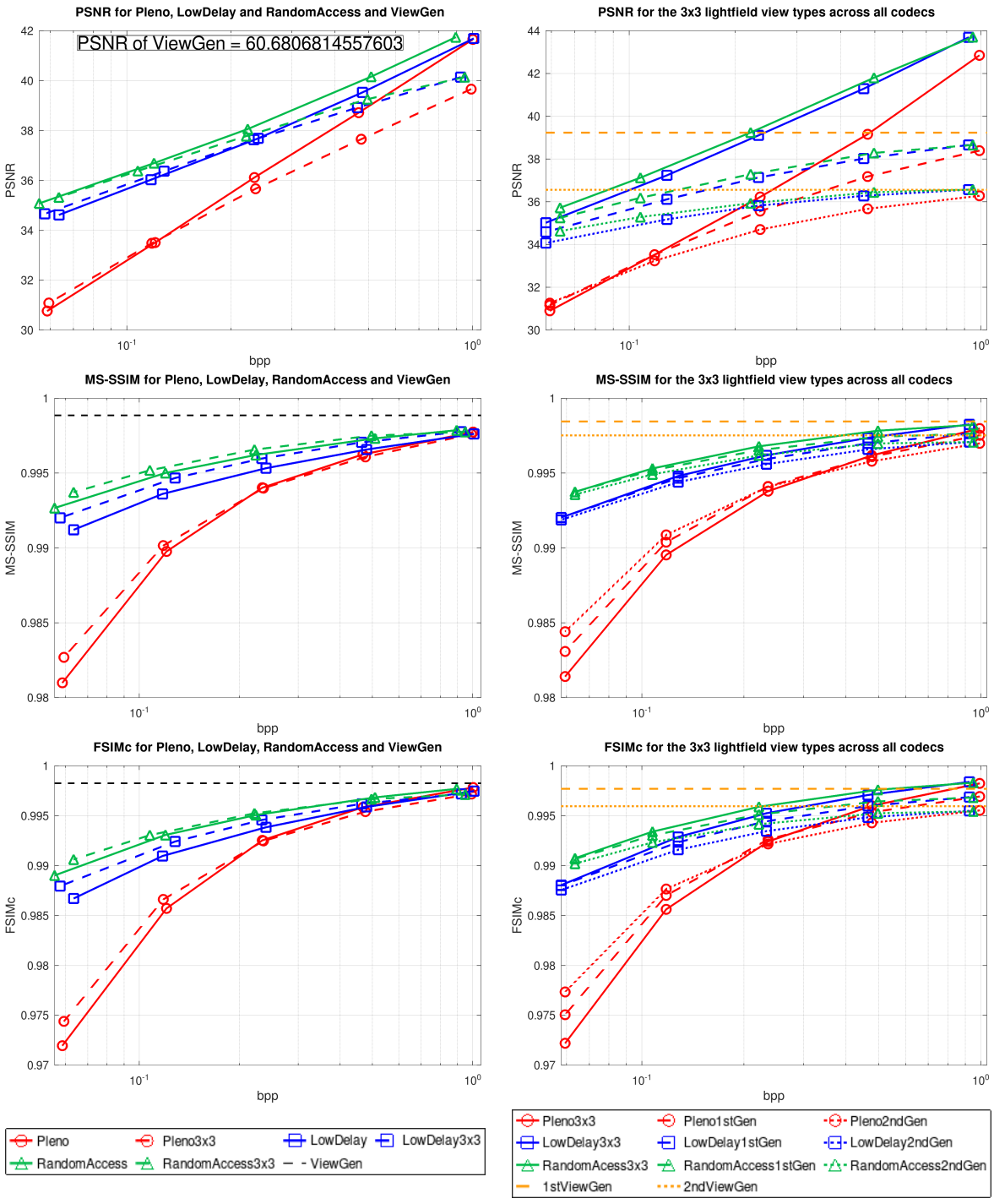


Table 4.2: Plots for the Fountain&Vincent 2 light field.

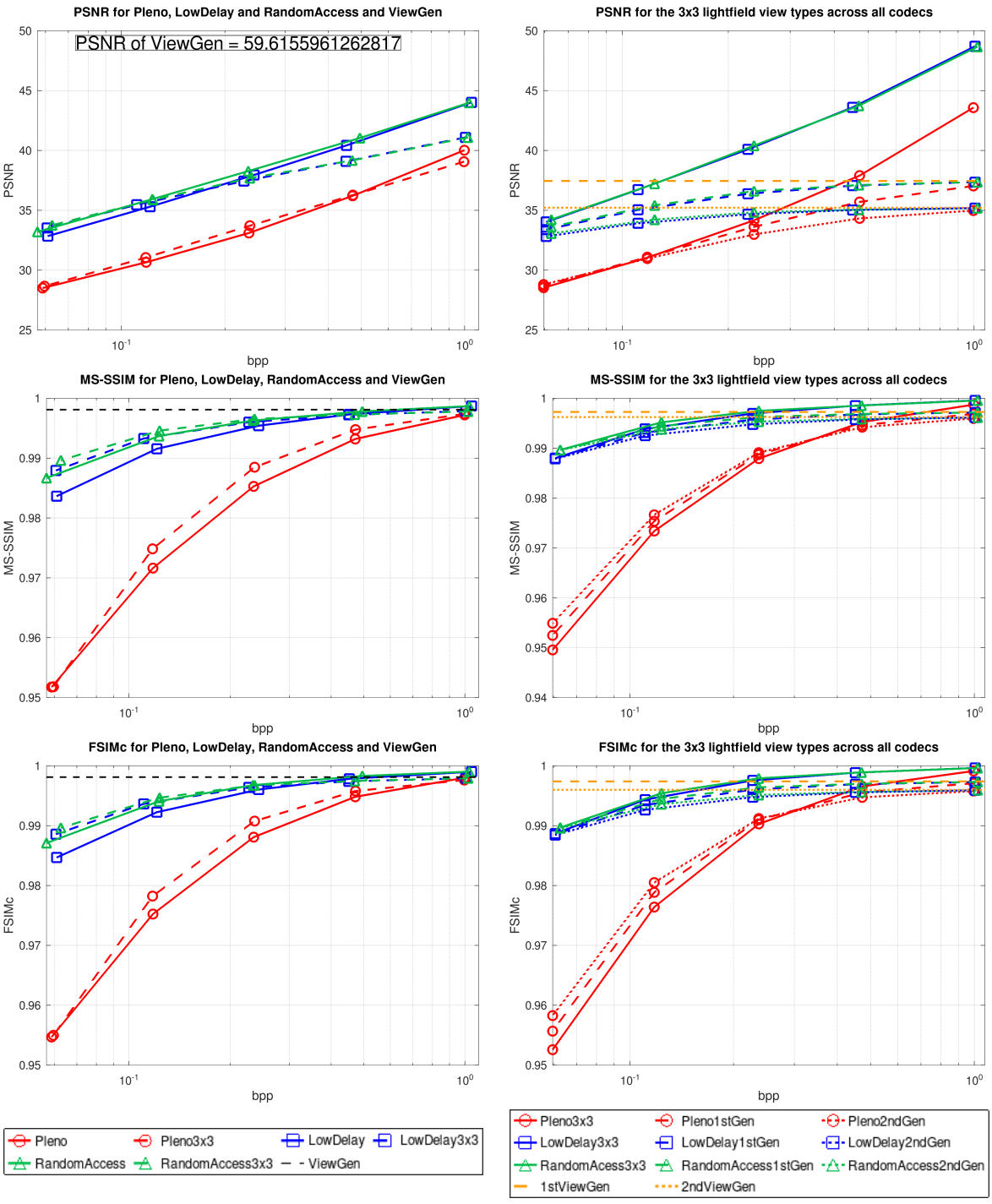


Table 4.3: Plots for the Bicyle light field.

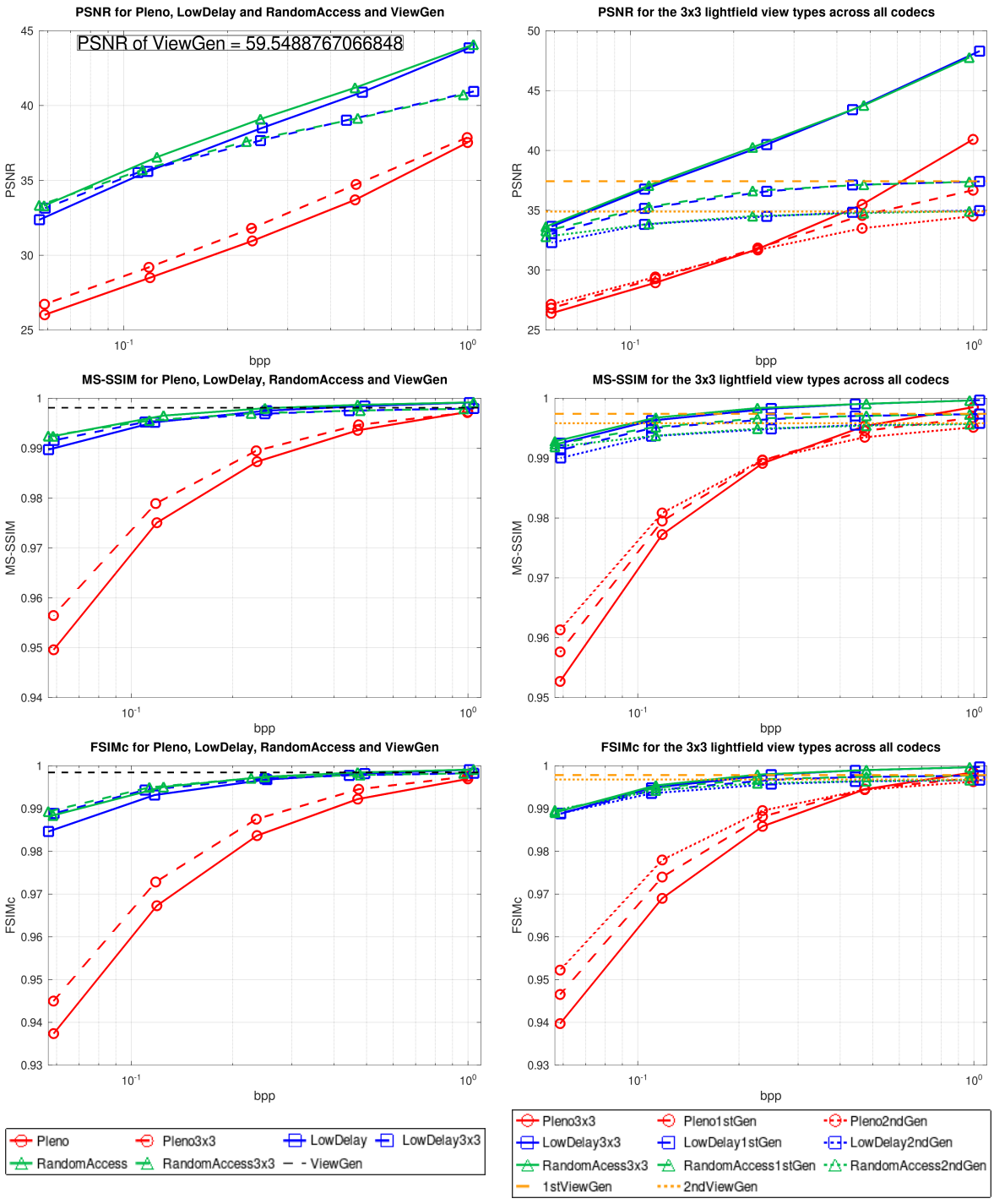


Table 4.4: Plots for the Sideboard light field.

4.1 Plots Analysis

4.1.1 Reference Line (ViewGen)

The black dashed reference line labeled “ViewGen” remains consistently high across all metrics, representing the performance of the view generation method without compression, providing a benchmark for comparison. Furthermore, by examining the yellow reference lines in the right column plots, it becomes easier to observe the effect of the view synthesis at each stage. While there is an expected loss of quality for each generation level, the initial drop in quality between the original 3×3 light field and the first generation views is more significant than the subsequent loss between the first and second generation views.

However, it is worth noting that the PSNR-HVS-M metric tends to be more affected compared to MS-SSIM and FSIMc. This difference can be attributed to the nature of the metrics: PSNR focuses on pixel-wise differences, making it more sensitive to even slight variations in pixel values, even with the modified version PSNR-HVS-M. In contrast, MS-SSIM and FSIMc are more perceptual metrics, designed to better align with human visual perception. They tend to be less impacted by small pixel-level changes, which is why they exhibit higher and more stable values in comparison to PSNR-HVS-M. This highlights how, even though PSNR-HVS-M shows a relatively larger drop, the actual perceived quality remains high across the other metrics.

This demonstrates the SepConv++ method ability to reconstruct the views with minimal quality loss, making it a reliable choice before introducing the additional compression losses.

4.1.2 Codec Performance (Pleno, VVC LowDelay, VVC Random Access)

Overall, Pleno consistently delivers the poorest performance across all metrics, with the gap being particularly noticeable at lower bitrates. VVC Random Access performs the best, with VVC Low Delay following closely but always slightly behind. However, as the bitrates increase, all three codecs converge, showing similar performance at higher bitrates.

4.1.3 Comparison Between Fully Compressed and View Synthesized Light Fields

The analysis is divided into two parts because PSNR-HVS-M and the perceptual metrics MS-SSIM and FSIMc evaluate different aspects of image quality. PSNR-HVS-M, as a pixel-based metric, measures pixel-level differences, while MS-SSIM and FSIMc focus on structural information and human visual perception. This leads to distinct behaviors

when assessing the plots generated for the different light fields.

4.1.3.1 PSNR-HVS-M plots behavior

For PSNR-HVS-M, the fully compressed 5×5 light field consistently outperforms the sparsely sampled 3×3 light field that undergoes view synthesis, across all datasets and bitrates. This performance gap becomes more pronounced at higher bitrates. There are only a few instances, primarily at lower bitrates, where the two sets exhibit comparable performance, but these are minimal and do not significantly impact the overall conclusions.

The behavior observed in the PSNR-HVS-M plot of the left columns can be further explained by analyzing the corresponding plots in the second columns, which breaks down the PSNR-HVS-M performance by light field view types. While all view types (original, first generation, and second generation) start with the same value at the lowest bitrate, the gap between them increases as the bitrate rises. Specifically, the original (3×3), the first generation synthesized views, and the second generation synthesized views drift further apart in terms of quality. This widening gap negatively impacts the overall performance of the 3×3 set, especially at higher bitrates, where the lower quality of the synthesized views becomes more evident, explaining why the 3×3 set performs worse compared to the fully compressed 5×5 set.

4.1.3.2 MS-SSIM and FSIMc plots behaviour

For VVC Random Access and VVC LowDelay: In the MS-SSIM and FSIMc plots, the sparsely sampled 3×3 light field that undergoes view synthesis, often performs better than the fully compressed 5×5 one, particularly at lower bitrates. As the bitrate increases, the gap between the two narrows, and in most cases, they converge at higher bitrates.

This behavior is further explained by analyzing the corresponding MS-SSIM and FSIMc plots for the view types in the second columns. For these cases, at the lowest bitrate, all view types exhibit nearly identical values for every codec. As the bitrate increases, the gap widens slightly, the original maintain the highest quality followed by first generation, and lastly second generation. The gap between these generations remains quite small, even at the highest bitrate where the difference is more pronounced.

This explains why, in most cases, the sparsely sampled 3×3 set performs comparably or better at lower bitrates. However, it gradually loses this advantage with the bitrate increase and the quality differences between view synthesis stages become more noticeable.

For Pleno:

The Pleno codec exhibits a different behavior when handling view generation, diverging from the patterns observed in the other codecs. By examining the second-column plots for all datasets, it becomes clear that at the lowest bitrate, the second-generation synthesized

views demonstrate the highest quality, followed by the first generation, and lastly, the original views. This is contrary to what is observed with other codecs, where the original views typically maintain the highest quality.

As the bitrate increases, the gap between these view types diminishes until they intersect. The exact point of intersection varies depending on the metric/light field, occurring at different bitrates—sometimes in the middle of the range and other times closer to the higher bitrates. After the intersection, the gap between the views widens again, but the order of quality reverses: original views now exhibit the best quality, followed by the first-generation views, and finally, the second-generation views.

This shift in behavior between view types explains why the first-column plots for Pleno show slight differences depending on the metric and dataset being analyzed. Although all second-column plots for MS-SSIM and FSIMc follow the same general behavior, the variations in values, such as gap sizes and intersection points, cause the left-column plots to exhibit subtle differences across the different datasets. These differences are listed below:

- **Bikes and Fountain Light Fields**: The sparsely sampled 3×3 light field that undergoes view synthesis, initially outperforms the fully compressed 5×5 light field at lower bitrates. However, as the bitrate increases, this gap gradually decreases until the two intersect from mid to high bitrates. After the intersection, the fully compressed 5×5 set surpasses the 3×3 that suffered view synthesis, although the difference remains minimal after the intersection.
- **Sideboard Light Field**: Similar to the Bikes and Fountain light fields, the sparsely sampled 3×3 light field that undergoes view synthesis starts with a better performance at the lowest bitrate compared to the fully compressed 5×5 one. As the bitrate increases, the gap narrows. For the highest bitrate, both sets converge, resulting in nearly identical performance.
- **Bicycle Light Field**: This dataset shows unique behavior. At the lowest bitrate, the 3×3 set that goes through view synthesis and the fully compressed 5×5 set begin with the same performance. From that point to the mid-bitrate range, the 3×3 set outperforms the 5×5 set, with the gap widening. However, at mid to high bitrates, the gap closes once again, and the two sets converge by the highest bitrate. This behavior could potentially be caused by the up-sampling for 10-bit depth.

4.2 Additional Results and Analysis

4.2.1 Bjontegaard Metrics

Table 4.5 shows the average Bjontegaard (BD) deltas for the PSNR-HVS-M, MS-SSIM and FSIMc considering the four light fields used in this work. For every codec the fully

Table 4.5: Average BD-Metrics and BD-Rate for each codec, comparing the 3×3 set against the 5×5 set.

Codec	PSNR-HVS-M		MS-SSIM		FSIMc	
	BD-PSNR-HVS-M	BD-Rate	BD-Metric	BD-Rate	BD-Metric	BD-Rate
Pleno	-0.024	0.515%	1.330E-03	-6.864%	1.544E-03	-3.958%
Low Delay	-0.797	25.341%	6.370E-05	-0.731%	3.762E-05	5.699%
Random Access	-0.425	9.494%	6.268E-04	-16.034%	7.605E-04	-12.154%

Table 4.6: Average BD-Metrics and BD-Rate considering the 5×5 JPEG Pleno as reference.

Codec set	PSNR-HVS-M		MS-SSIM		FSIMc	
	BD-PSNR-HVS-M	BD-Rate	BD-Metric	BD-Rate	BD-Metric	BD-Rate
Random Access5x5	4.211	-61.184%	8.812E-03	-65.866%	9.599E-03	-59.712%
LowDelay5x5	3.690	-54.791%	7.775E-03	-55.084%	8.340E-03	-48.895%
Random Access3x3	3.413	-56.796%	8.798E-03	-68.404%	9.548E-03	-59.229%
LowDelay3x3	3.312	-55.308%	8.508E-03	-63.917%	9.244E-03	-56.547%

compressed light fields (5×5) is used as reference to be compared with the correspondent sparsely sampled light field (3×3) that undergoes view synthesis. It can be observed that a better PSNR-HVS-M is obtained using fully compressed light fields for every codec instead of a small set of views followed by view synthesis. However, when perceptual metrics such as MS-SSIM and FSIMc are used, a slight increase in compression performance is obtained using the selection of 3×3 and the view synthesis method.

Table 4.6 presents a comparison using Bjontegaard metrics of the Pleno codec's fully compressed 5×5 set with various other codec configurations. For PSNR-HVS-M, the results show that all other codecs outperform the Pleno 5×5 set. The ranking, from best to worst quality, is as follows: Random Access 5×5, LowDelay 5×5, Random Access 3×3, LowDelay 3×3, Pleno 5×5, and Pleno 3×3.

For perceptual metrics like MS-SSIM and FSIMc, a similar trend is observed, where all light fields surpass JPEG Pleno 5×5 light field. In this case, the quality ranking from best to worst is: Random Access 5×5, Random Access 3×3, LowDelay 3×3, LowDelay 5×5, Pleno 3×3, and finally, Pleno 5×5.

4.2.2 Visual Comparison

In this section, a visual comparison of the reconstructed light fields is presented to complement the metric-based analysis. Since human perception of image quality does not always align with objective metrics, these visual comparisons offer insight into how compression and view synthesis affect the reconstructed views. Several figures are provided to show the results across different datasets and codec configurations.

Figure 4.1 focuses on the Bikes dataset, compressed using JPEG Pleno. Its organized as a 3x6 matrix, where the first row displays three original views: the first is part of the 3×3 set, while the second and third are dropped views, later reconstructed via first and

second-stage view synthesis. Below each original view, the subsequent rows display the compressed or generated versions at five different bitrates: very high, high, medium, low, and very low, listed from top to bottom. Similarly, Figure 4.3 provides visual comparisons for the Fountain and Vincent light field using the same structure. For the Sideboard (Figure 4.5) and Bicycle (Figure 4.8) light fields, only three rows—medium, low, and very low bitrates—are displayed due to larger image sizes and minimal visual differences at higher bitrates.

For the VVC codec configurations (Low Delay and Random Access), only two bitrate values — medium and very low — are presented in the visual evaluations. This is because VVC performs so well perceptually, that very high and high bitrates do not show noticeable differences in quality, even when compared to the original views. In fact, even at very low bitrates, the compression quality remains visually impressive, reducing the need to include additional higher bitrate comparisons.

The analysis for Low Delay and Random Access codecs follows a similar structure. Figure 4.2, for the Bikes light field, begins with the original views in the first row, followed by two rows displaying the compressed versions at medium and very low bitrates using Low Delay. After a repeated row of original views, the next two rows show the results for Random Access at the same bitrates. Figure 4.4 uses the same layout for the Fountain&Vincent light field. Due to the larger image sizes of the Sideboard light field, the results are split between two figures: Figure 4.6 shows the first three rows for Low Delay, and Figure 4.7 shows the corresponding Random Access rows. This division is also used for the Bicycle dataset, with Figures 4.9 and 4.10 showing Low Delay and Random Access results, respectively.

The views selected for analysis were determined by using the JPEG Pleno codec as a reference for each light field. The view with the lowest MS-SSIM was chosen from the highest bitrate across each view type: original compressed views, first-generation views, and second-generation views.

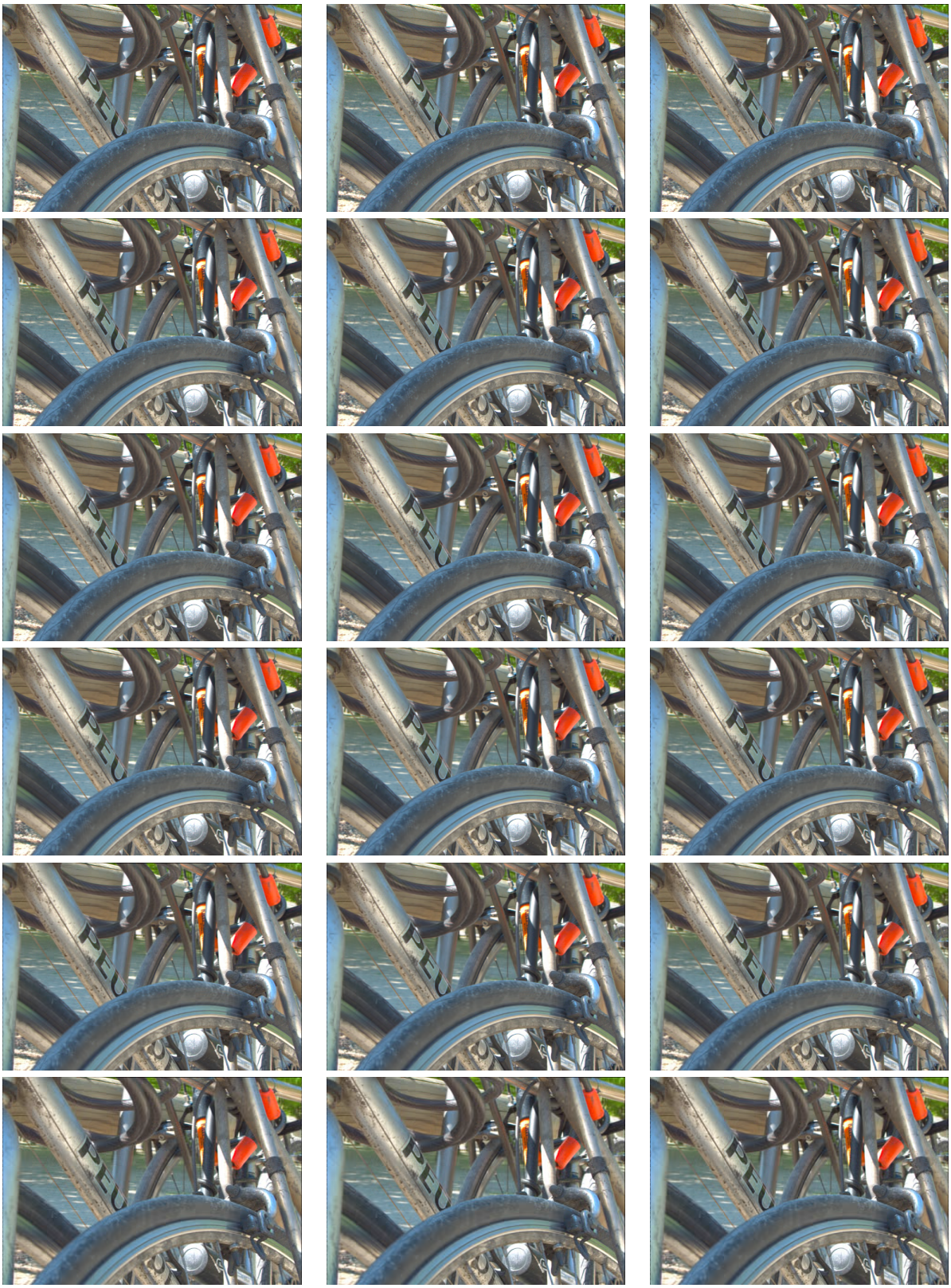


Figure 4.1: Chosen views for visual evaluation of JPEG Pleno for Bikes Light Field.

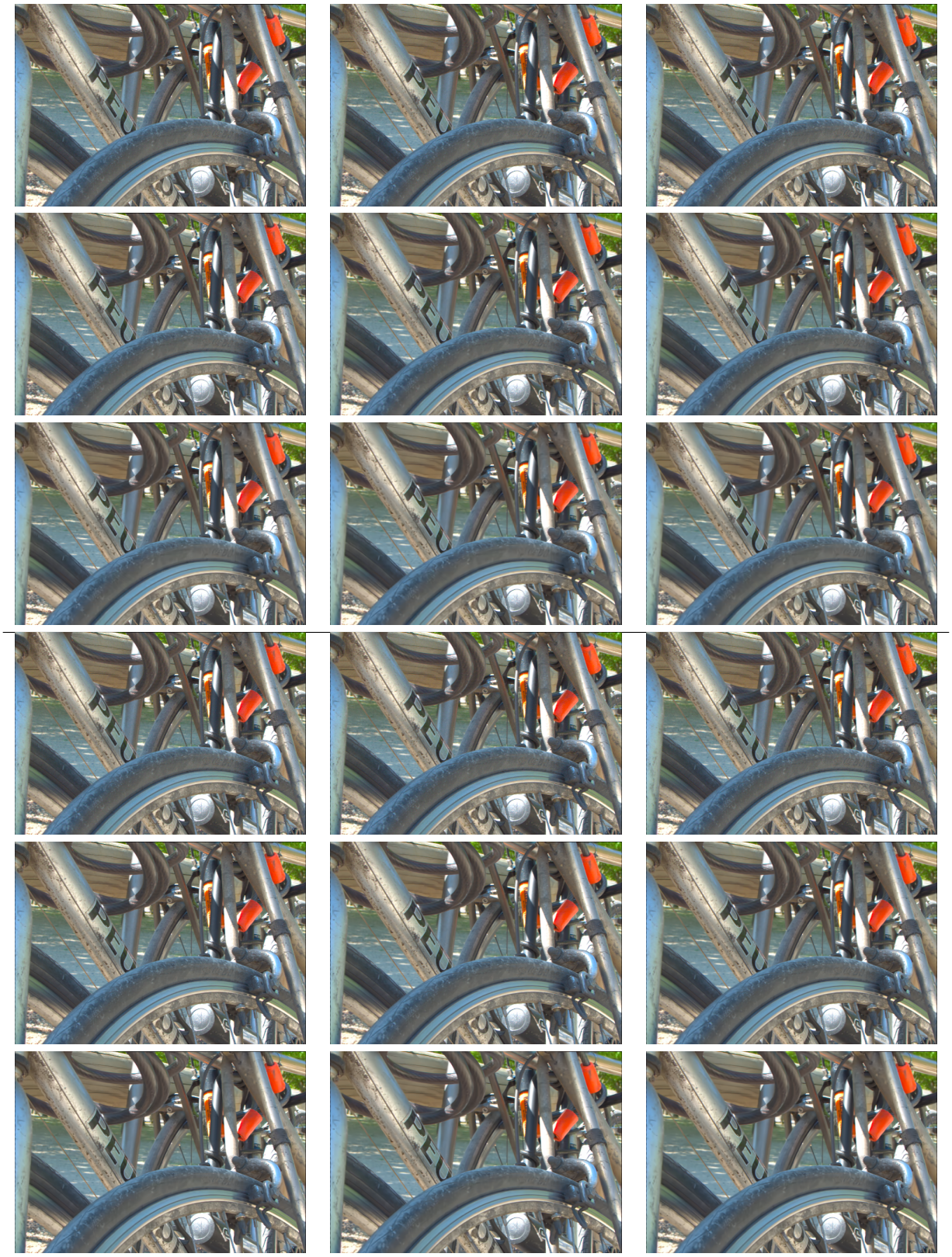


Figure 4.2: Chosen views for visual evaluation of VVCs Low Delay and Random Access configurations for Bikes Light Field.

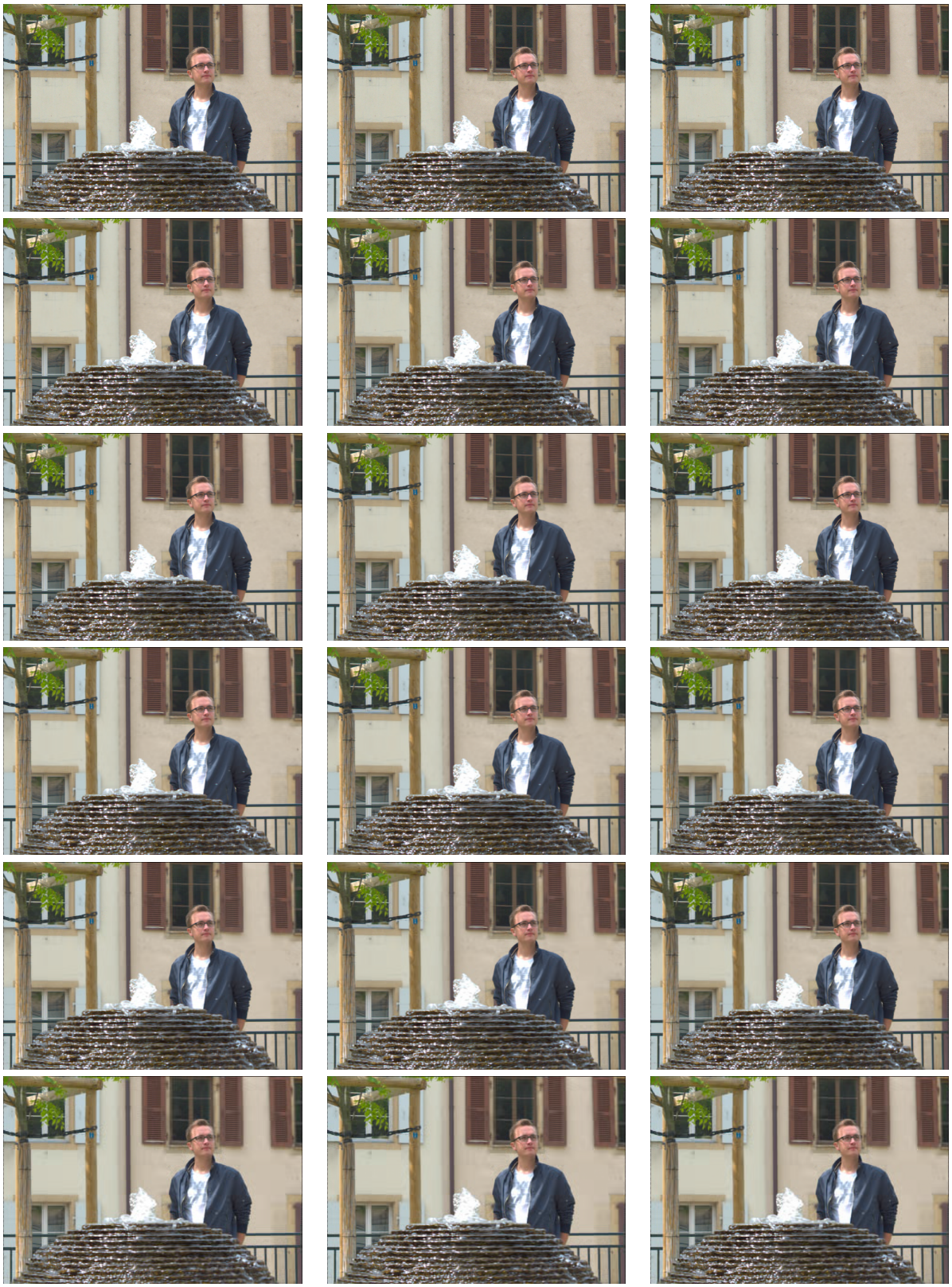


Figure 4.3: Chosen views for visual evaluation of JPEG Pleno for Fountain&Vincent 2 Light Field.

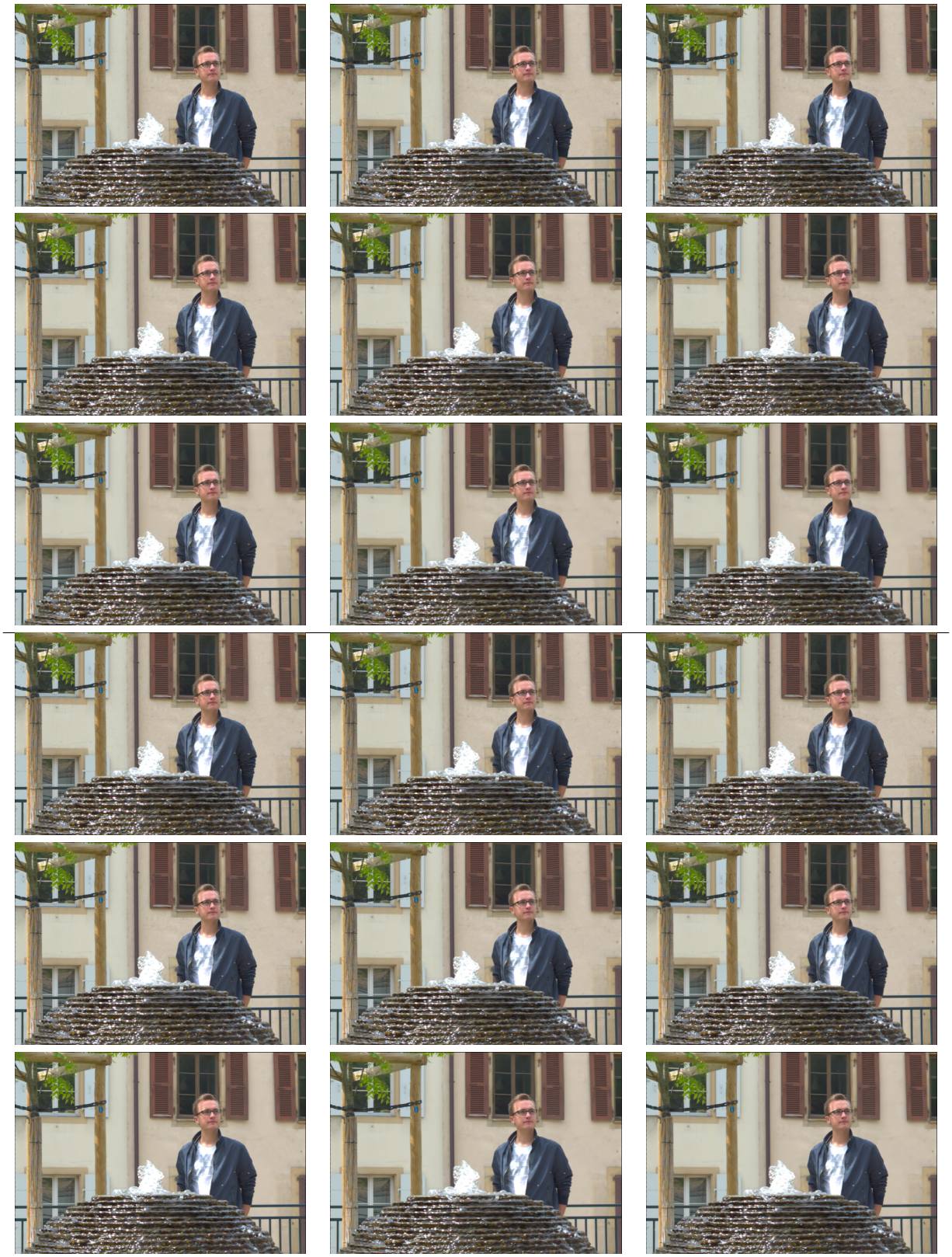


Figure 4.4: Chosen views for visual evaluation of VVCs Low Delay and Random Access configurations for Fountain&Vincent 2 Light Field.

Figure 4.5: Chosen views for visual evaluation of JPEG Pleno for Sideboard Light Field

Figure 4.6: Chosen views for visual evaluation of VVC with Low Delay configuration for Sideboard Light Field.

Figure 4.7: Chosen views for visual evaluation of VVC with Random Access configuration for Sideboard Light Field.

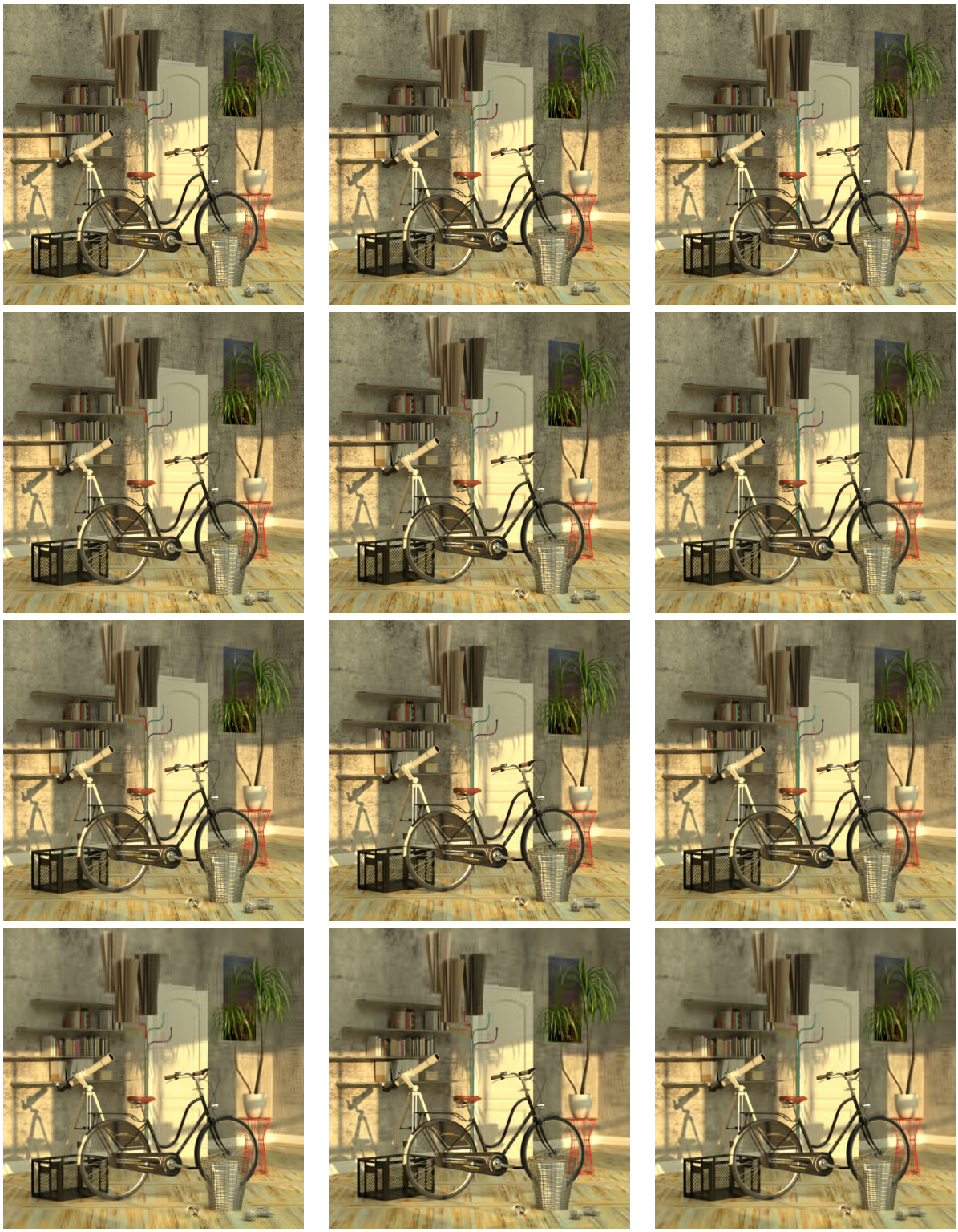


Figure 4.8: Chosen views for visual evaluation of JPEG Pleno for Bicycle Light Field.

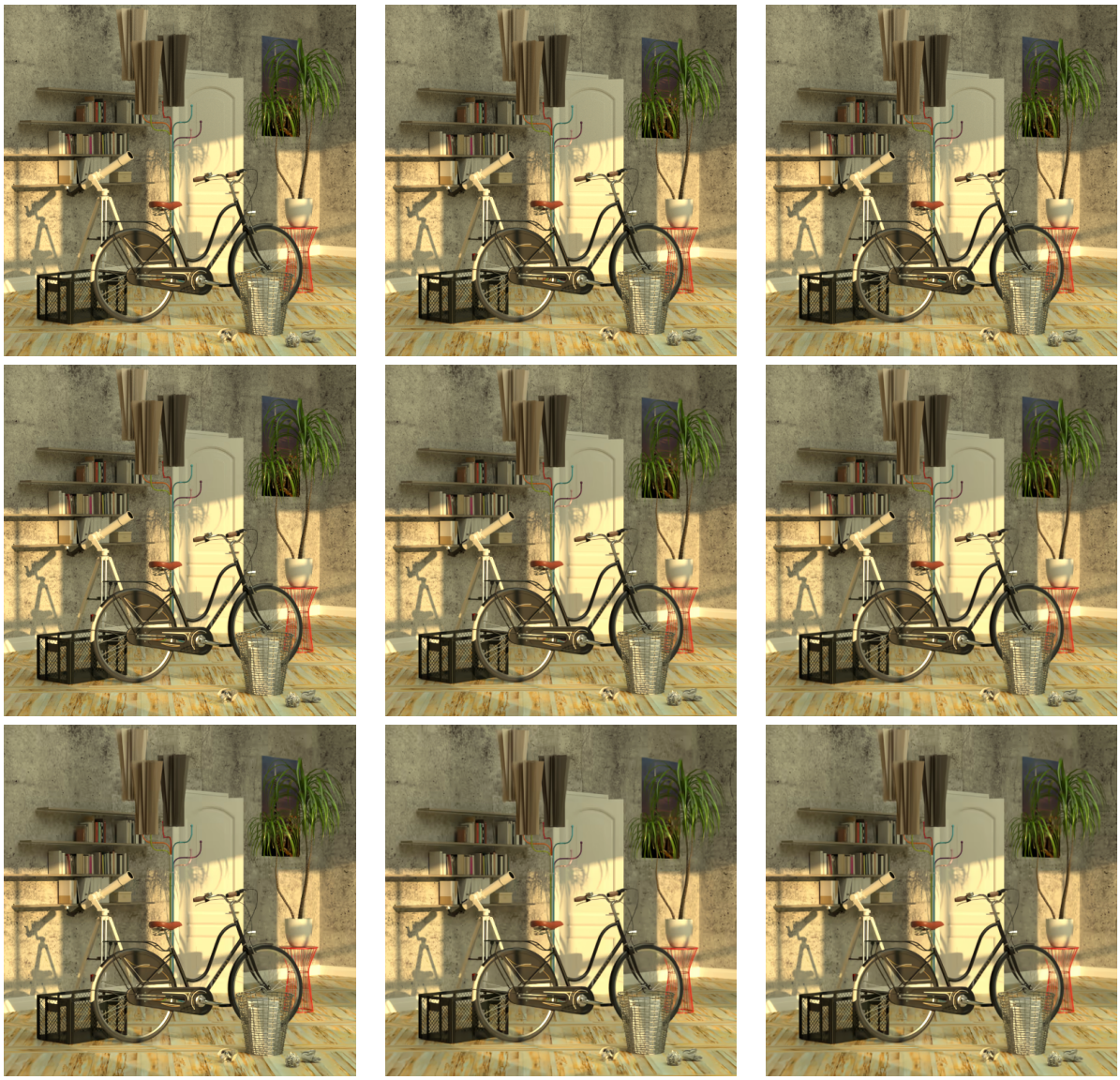


Figure 4.9: Chosen views for visual evaluation of VVC with Low Delay configuration for Bicycle Light Field

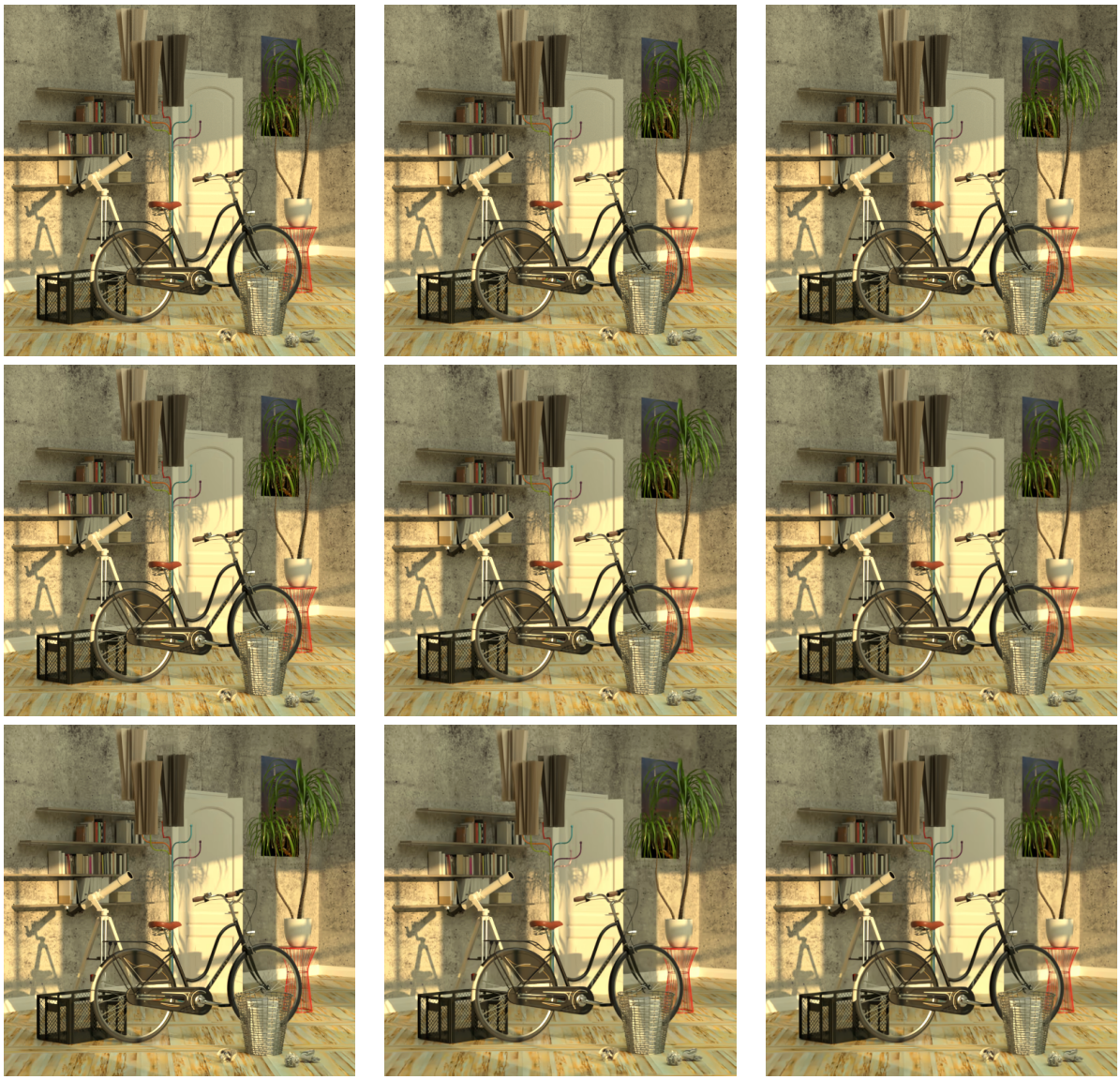


Figure 4.10: Chosen views for visual evaluation of VVC with Random Access configuration for Bicycle Light Field

4.2.3 Compression Times

The compression times were measured on a system running Ubuntu 22.04.5 LTS with an AMD Ryzen 7 2700X Eight-Core Processor and 32 GB of RAM. The times shown in the table below represent the average compression time (in seconds) for each codec and bitrate, calculated across four different datasets. Results are provided for both the fully compressed 5×5 set and the sparsely sampled 3×3 set.

	Pleno	**Pleno3x3**	**Low Delay**	**Low Delay 3x3**	**Random Access**	**Random Access 3x3**
Very High	12,20	8,31	2536,6	1376,9	3209,7	1833,3
High	8,69	6,00	1847,1	951,7	2033,6	1225,411
Medium	6,748	4,113	1280,983	660,509	1281,618	794,346
Low	5,452	3,041	863,684	448,303	790,970	493,899
Very Low	4,517	2,352	521,578	282,181	478,711	318,799

Table 4.7: Compression times across different datasets for various codecs and configurations

1. **Bitrate and Encoding Time Relationship:**
   - As the bitrate decreases from Very High to Very Low, encoding times consistently decrease across all codecs. This reinforces the idea that lower bitrates require less computational effort, aligning with expectations in image compression.
2. **3×3 Sparse Sampling Efficiency:**
   - The 3×3 sparsely sampled light fields (Pleno3×3, Low Delay 3×3, Random Access 3×3) show notably lower encoding times compared to their fully compressed 5×5 counterparts. Though by adding the view synthesis process time no substancial difference will be noticed for the VVC codec (in either configuration), since JPEG Pleno is substantially quicker the addition of the view synthesis process would actually make Pleno3×3 slower than the fully compressed Pleno 5×5

.

3. **Comparative Performance of VVC vs. JPEG Pleno:**
   - While VVC codec (Low Delay and Random Access configurations) generally provide better performance than JPEG Pleno across most bitrate categories, the encoding time for VVC is much higher. For example, in the Very High bitrate category, the encoding time for VVC (3209.7 seconds for Random Access) starkly contrasts with JPEG Pleno (12.20 seconds). This raises a critical consideration: the performance gains might not be justified by the additional computational effort required, especially in time-sensitive applications.

Chapter 5

Conclusions and Future Work

This thesis aimed to explore the potential of view synthesis to improve light field compression while reducing the complexity, and focusing on maintaining high image quality, while reducing data storage and transmission requirements. The study examined two codecs, one of them using two distinct configurations — JPEG Pleno, VVC Low Delay, and VVC Random Access — across datasets of distinct characteristics to evaluate their performance alongside view synthesis. By applying compression to both fully captured and sparsely sampled light fields, the work aimed to determine how effectively fewer views could be used while still reconstructing the original light field with minimal quality degradation.

This work allows to conclude that the SepConv++ can use compressed views to generate denser light fields without a relevant loss of quality. For lower qualities, the perceptual metrics even show slightly better quality.

Sparse light fields require less bitrate, but to obtain similar average qualities after view synthesis, a very similar bit rate is needed.

The use of the proposed model has the advantage of requiring much less complexity and resources, because only a small set of views needs to be compressed, while other views can be synthesized with similar quality levels.

VVC consistently outperforms JPEG Pleno in retaining quality, whether using Random Access or Low Delay configurations. Despite this, JPEG Pleno is significantly faster in execution, offering a relevant lower complexity advantage.

Additionally, VVC supports both 8-bit and 10-bit depth images, making it more versatile. In contrast, the current JPEG Pleno implementation is limited to 10-bit images, which limits its practicality for other types of images, as they require additional processing before use.

This study explored the potential of learning-based view synthesis techniques, particularly focusing on the SepConv++ model, for light field compression. The results demonstrate that SepConv++ can successfully generate denser light fields from compressed views without a significant loss in quality. Notably, for higher quality reconstructions, perceptual metrics such as MS-SSIM and FSIMc even indicated a slight improvement, emphasizing the effectiveness of the view synthesis process.

The analysis also highlighted the balance between compression efficiency and reconstruction quality. Sparse light fields require lower bitrate for storage and transmission, when the parameters used for compression are the same as the ones used for the fully com-

pressed light field. However, to achieve similar average quality levels after view synthesis, an increase in bitrate to nearly similar bitrates values used for the fully compressed light fields, is necessary.

From a codec comparison perspective, VVC consistently outperformed JPEG Pleno in retaining visual quality across both Random Access and Low Delay configurations. However, JPEG Pleno offered a notable speed advantage, making it a viable option for applications where fast encoding is a priority.

The proposed approach, which utilizes compressed views for synthesis, proved advantageous by significantly reducing computational complexity and resource demands, particularly evident in the VVC codec, where the encoding time is substantially higher for fully compressed light fields. Only a subset of the views needed to be compressed and transmitted, while the remaining views were synthesized with minimal quality loss, making this a promising method for light field data handling. However, the initial premise that view synthesis would allow for a reduced bitrate while maintaining the same quality was not fully supported by the results. In practice, with the selected tools, it was observed that achieving comparable quality levels required bitrates that were nearly the same as those for fully compressed light fields. This revealed how effective this codecs are in extracting the redundancies between views.

5.1 Future Work

Several avenues for future research can be pursued based on the findings of this study:

- **Advanced View Synthesis Models**: Further exploration of more sophisticated new view synthesis techniques, including those based on deep learning models, could potentially enhance the quality of synthesized views.
- **Sub-sampling Strategies**: Further investigation into different sub-sampling strategies and ratios could reveal optimal configurations for various light field datasets. Recent studies utilizing view synthesis for super-resolution purposes have demonstrated the effectiveness of methods where only corner input views are required to generate the rest of the light field, which could significantly reduce the amount of data that needs to be stored.
- **Broader Dataset Evaluation**: Extending the evaluation to a more diverse range of datasets would allow for a better understanding of the performance and flexibility of the proposed methods across different light field scenarios.
- **Subjective Testing**: Due to time constraints, it was no possible to include a subjective evaluation of the results. However, conducting such tests would be essential to better understand the perceptual quality of the compressed and synthesized

views, as subjective assessments often provide insights that objective metrics may overlook.

In conclusion, the use of view synthesis in light field compression represents a promising approach to reduce storage and transmission costs while maintaining high visual quality. However, further advancements in synthesis techniques, codec implementations, and sub-sampling strategies are necessary to fully realize the potential of this technology.

Bibliography

[1] S. Zhou, T. Zhu, K. Shi, Y. Li, W. Zheng, and J. Yong, "Review of light field technologies," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, p. 29, Dec. 2021. xvii, 6, 10

[2] "The Stanford Multi-Camera Array." `https://graphics.stanford.edu/projects/array/`. [Accessed Oct-2024]. xvii, 7

[3] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light Field Photography with a Hand-Held Plenopic Camera," *Technical Report CTSR 2005-02*, vol. CTSR, 01 2005. xvii, 7

[4] Z. Hossain, A. S. Backer, and Y. Chen, "CS348b Project: Light Field Camera Simulation," xvii, 8

[5] W. Feng, J. Gao, T. Qu, S. Zhou, and D. Zhao, "Three-dimensional reconstruction of light field based on phase similarity," *Sensors*, vol. 21, no. 22, 2021. xvii, 9

[6] N. Balram and I. Tošić, "Light□Field Imaging and Display Systems," *Information Display*, vol. 32, pp. 6–13, 08 2016. xvii, 6, 8, 9

[7] W. Fu, X. Tong, C. Shan, S. Zhu, and B. Chen, "Implementing light field image refocusing algorithm," in *2015 2nd International Conference on Opto-Electronics and Applied Optics (IEM OPTRONIX)*, pp. 1–8, 2015. xvii, 10

[8] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, "DeOccNet: Learning to See Through Foreground Occlusions in Light Fields," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. xvii, 11

[9] V. Vaish, B. Wilburn, N. Joshi, and M. Levoy, "Using plane + parallax for calibrating dense camera arrays," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, 2004. xvii, 11

[10] Z. Pei, X. Chen, and Y.-H. Yang, "All-In-Focus Synthetic Aperture Imaging Using Image Matting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 2, pp. 288–301, 2018. xvii, 11

[11] G. De Oliveira Alves, M. B. De Carvalho, C. L. Pagliari, P. G. Freitas, I. Seidel, M. P. Pereira, C. F. S. Vieira, V. Testoni, F. Pereira, and E. A. B. Da Silva, "The JPEG Pleno Light Field Coding Standard 4D-Transform Mode: How to Design an Efficient 4D-Native Codec," *IEEE Access*, vol. 8, pp. 170807–170829, 2020. xvii, 15

[12] M. B. d. Carvalho, C. L. Pagliari, G. d. O. E. Alves, C. Schretter, P. Schelkens, F. Pereira, and E. A. B. d. Silva, "Supporting Wider Baseline Light Fields in JPEG

Pleno With a Novel Slanted 4D-DCT Coding Mode," *IEEE Access*, vol. 11, pp. 28294–28317, 2023. xvii, 14, 15

[13] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021. xvii, 16

[14] S. Niklaus, L. Mai, and F. Liu, "Video Frame Interpolation via Adaptive Separable Convolution," in *IEEE International Conference on Computer Vision*, 2017. xvii, 20, 21

[15] S. Niklaus, L. Mai, and O. Wang, "Revisiting Adaptive Convolutions for Video Frame Interpolation," in *IEEE Winter Conference on Applications of Computer Vision*, 2021. xvii, 20, 21

[16] A. Gershun, "The Light Field," *Journal of Mathematics and Physics*, vol. 18, no. 1-4, pp. 51–151, 1939. 5

[17] M. Levoy and P. Hanrahan, "Light Field Rendering," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, (New York, NY, USA), Association for Computing Machinery, 2023. 5, 14

[18] L. Yu, Y. Ma, S. Hong, and K. Chen, "Reivew of Light Field Image Super-Resolution," *Electronics*, vol. 11, p. 1904, Jan. 2022. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute. 6, 8, 11

[19] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light Field Image Processing: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017. 7, 20

[20] G. Wetzstein, I. Ihrke, D. Lanman, W. Heidrich, R. Raskar, and K. Akeley, "Computational plenoptic imaging," in *ACM SIGGRAPH 2012 Courses*, SIGGRAPH '12, (New York, NY, USA), Association for Computing Machinery, 2012. 7

[21] M. Baradad, V. Ye, A. B. Yedidia, F. Durand, W. T. Freeman, G. W. Wornell, and A. Torralba, "Inferring light fields from shadows," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6267–6275, 2018. 10

[22] H. Sheng, P. Zhao, S. Zhang, J. Zhang, and D. Yang, "Occlusion-aware depth estimation for light field using multi-orientation epis," *Pattern Recognition*, vol. 74, pp. 587–599, 2018. 10

[23] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12095–12103, Apr. 2020. 10

[24] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4748–4757, 2018. 10

[25] S. Wanner, C. Straehle, and B. Goldluecke, "Globally Consistent Multi-label Assignment on the Ray Space of 4D Light Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 11

[26] D. R. Horn and B. Chen, "Lightshop: interactive light field manipulation and rendering," in *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*, I3D '07, (New York, NY, USA), p. 121–128, Association for Computing Machinery, 2007. 11

[27] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to Deblur using Light Field Generated and Real Defocus Images," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16283–16292, 2022. 11

[28] S. Mahmoudpour, C. Pagliari, and P. Schelkens, "Learning-based light field imaging: an overview," *EURASIP Journal on Image and Video Processing*, vol. 2024, 05 2024. 12

[29] K. Ko, Y. J. Koh, S. Chang, and C.-S. Kim, "Light field super-resolution via adaptive feature remixing," *IEEE Transactions on Image Processing*, vol. 30, pp. 4114–4128, 2021. 12

[30] G. Wu, Y. Wang, Y. Liu, L. Fang, and T. Chai, "Spatial-angular attention network for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 30, pp. 8999–9013, 2021. 12

[31] "Lytro company fact sheet." `http://www.lytro.com/lytro_company_factsheet.pdf`, 2012. [Accessed Oct-2024]. 13

[32] Raytrix, "Raytrix light field camera." `http://www.raytrix.de/`, 2024. [Accessed Oct-2024]. 13

[33] Wooptix, "The ultimate image solutions." `https://wooptix.com/`, 2021. [Accessed Oct-2024]. 13

[34] C. Pitts, C.-K. Liang, and K. Akeley, "Capturing light-field images with uneven and/or incomplete angular sampling," July 24 2018. US Patent 10,033,986. 13

[35] S. Corporation, "About spatial reality display." `https://www.sony.net/Products/Developer-Spatial-Reality-display/en/develop/AboutSRDisplay.html`, 2024. [Accessed Oct-2024]. 13

[36] S. I. Studios, "Sony pictures." `https://www.sonyinnovationstudios.com/`, 2024. [Accessed Oct-2024]. 13

[37] L. G. Factory, "Looking glass factory." https://lookingglassfactory.com/, 2024. [Accessed Oct-2024]. 13

[38] Google, "Project starline: Feel like you're there, together." https://blog.google/technology/research/project-starline/, 2024. [Accessed Oct-2024]. 13

[39] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, "JPEG Pleno: Toward an Efficient Representation of Visual Reality," *IEEE MultiMedia*, vol. 23, no. 4, pp. 14–20, 2016. 14

[40] "Information Technology Plenoptic Image Coding System (JPEG Pleno) – Part 2: Light Field Coding, ISO/IEC 21794-2:2021," tech. rep., ISO/IEC, Geneva, Switzerland, apr 2021. 14

[41] "Cd 15444-16 3ED (JPEG 2000 Part 16)," ,ISO/IEC JTC 1/SC29/WG1 N100805, ICQ, 103rd Meeting, Online, April 2024. 14

[42] "Information Technology — Coded Representation of Immersive Media — Part 3: Versatile Video Coding," Tech. Rep. N 18692, ISO/IEC JTC 1/SC 29/WG 11, July 2019. 15

[43] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012. 15

[44] M. Řeřábek and T. Ebrahimi, "New Light Field Image Dataset," in *8th International Workshop on Quality of Multimedia Experience (QoMEX)*, (Lisbon, Portugal), 2016. 18

[45] M. Řeřábek and T. Ebrahimi, "EPFL Light-Field Image Dataset." http://mmspg.epfl.ch/EPFL-light-field-image-dataset, 2016. Accessed: [insert date of access]. 18

[46] "Common Test Conditions for JPEG Pleno Light Field Quality Assessment v1.3." ISO/IEC JTC 1/SC 29/WG 1 (ITU-T SG16), 101st Meeting – Online – October 30-November 3, 2023, 2023. 18

[47] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields," in *Computer Vision – ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III*, (Berlin, Heidelberg), p. 19–34, Springer-Verlag, 2017. 18

[48] C. Perra, P. G. Freitas, I. Seidel, and P. Schelkens, "An overview of the emerging JPEG Pleno standard, conformance testing and reference software," in *Optics, Photonics and Digital Technologies for Imaging Applications VI* (P. Schelkens and

T. Kozacki, eds.), vol. 11353, pp. 207 – 219, International Society for Optics and Photonics, SPIE, 2020. 19

[49] "Vtm reference software for vvc." https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. [Accessed Oct-2024]. 19

[50] Y. Chen et al., "A Study of Efficient Light Field Subsampling and Reconstruction Strategies," *arXiv*, 2020. 20

[51] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light Field Reconstruction Using Shearlet Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 133–147, 2018. 20

[52] "JPEG AI Common Training and Test Conditions." ISO/IEC JTC 1/SC 29/WG 1 (ITU-T SG16), 98th Meeting – Sidney – January 16-20, 2023. 25, 26

[53] "JPEG AI Quality Assessment Framework." https://gitlab.com/wg1/jpeg-ai/jpeg-ai-qaf. [Accessed Oct-2024]. 25

[54] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008. 25

[55] K. Egiazarian, J. Astola, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," CD-ROM Proceedings of the Second International Workshop on Video Processing and Quality Metrics, Scottsdale, USA, 2006, 4 p. 25

[56] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, pp. 305–311, 2011. 25

[57] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proceedings of the third international workshop on video processing and quality metrics*, vol. 4, Scottsdale USA, 2007. 25

[58] "Psnr-hvs-m download page." https://www.ponomarenko.info/psnrhvsm.htm. [Accessed Oct-2024]. 25

[59] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," vol. 2, pp. 1398 – 1402 Vol.2, 12 2003. 25, 26

[60] D. M. Rouse and S. S. Hemami, "Understanding and simplifying the structural similarity metric," in *2008 15th IEEE International Conference on Image Processing*, pp. 1188–1191, Oct 2008. 26

[61] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A Feature Similarity Index for Image Quality Assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. 26

Appendix A

Appendix

A.1 Additional Data of the Light Fields

	Bitstream size	BPP	Control Parameter
PLENO	6764512	0.998	200
	3201352	0.472	599
	1597136	0.236	1600
	800368	0.118	5500
	401992	0.059	20000
PLENO 3x3	844991	0.997	77
	399955	0.472	240
	199471	0.235	695
	100070	0.118	1970
	50258	0.059	6100
LowDelay	6115736	0.902	17
	3150088	0.465	20
	1606072	0.237	23
	821168	0.121	26
	380392	0.056	30
LowDelay 3x3	700699	0.827	15
	375976	0.444	18
	204575	0.241	21
	105676	0.125	24
	50178	0.059	28
RandomAccess	912128	1.076	16
	406396	0.479	19
	215252	0.254	22
	94756	0.112	26
	54690	0.065	29
Random Access 3x3	952032	1.123	13
	368802	0.435	17
	208347	0.246	20
	101739	0.120	24
	50325	0.059	28

Table A.1: Bitstream Size (in bytes), bitrate (BPP) and bitrate control parameter defined for the Bikes light field.

	Bitstream Size	BPP	Control Parameter
PLENO	850480	1.003	230
	397354	0.469	750
	198266	0.234	2100
	102310	0.121	6000
	49825	0.059	22000
PLENO 3x3	841002	0.992	77
	404164	0.477	300
	199834	0.236	880
	100047	0.118	2600
	50358	0.059	7400
LowDelay	855165	1.009	17
	407555	0.481	21
	203455	0.240	24
	99522	0.117	27
	53881	0.064	30
LowDelay 3x3	784190	0.925	14
	394027	0.465	18
	197727	0.233	22
	108516	0.128	25
	49034	0.058	29
RandomAccess	758923	0.895	17
	431786	0.509	20
	189618	0.224	24
	101534	0.120	27
	47211	0.056	31
RandomAccess 3x3	805212	0.950	14
	421412	0.497	17
	187769	0.222	22
	91137	0.108	26
	53847	0.064	29

Table A.2: Bitstream size (in bytes), bitrate (BPP) and bitrate control parameter defined for the Fountain&Vincent light field.

	Bitstream Size	BPP	Control Parameter
PLENO	817800	0.998	180
	386382	0.472	720
	192658	0.235	2685
	96635	0.118	8800
	48257	0.059	26000
PLENO 3x3	815365	0.995	60
	387055	0.472	300
	193865	0.237	900
	96386	0.118	2840
	48860	0.060	10000
LowDelay	857524	1.047	13
	371023	0.453	18
	199337	0.243	22
	99251	0.121	26
	49956	0.061	30
LowDelay 3x3	824137	1.006	8
	369772	0.451	14
	186347	0.227	19
	90739	0.111	24
	49717	0.061	28
RandomAccess	843644	1.030	13
	405555	0.495	17
	191835	0.234	22
	100757	0.123	26
	46593	0.057	31
RandomAccess 3x3	837059	1.022	8
	385193	0.470	14
	193805	0.237	19
	101135	0.123	24
	51411	0.063	29

Table A.3: Bitstream size (in bytes), bitrate (BPP) and bitrate control parameter defined for the Bicycle light field.

	Bitstream Size	BPP	Control Parameter
PLENO	818987	1.000	465
	385558	0.471	1750
	193661	0.236	5000
	97587	0.119	13000
	48177	0.059	44500
PLENO 3x3	816817	0.997	133
	388072	0.474	610
	192589	0.235	2000
	96848	0.118	5800
	48122	0.059	16000
LowDelay	827491	1.010	13
	404635	0.494	17
	207098	0.253	21
	96104	0.117	26
	46510	0.057	31
LowDelay 3x3	853359	1.042	7
	363788	0.444	14
	204516	0.250	18
	90064	0.110	24
	48330	0.059	29
RandomAccess	851817	1.040	13
	385260	0.470	17
	204378	0.249	21
	102176	0.125	26
	47979	0.059	32
RandomAccess 3x3	796204	0.972	8
	392404	0.479	14
	186025	0.227	19
	92726	0.113	25
	46425	0.057	31

Table A.4: Bitstream size (in bytes), bitrate (BPP) and bitrate control parameter defined for the Sideboard light field.