Estimación de la profundidad a partir de proyecciones 2D del campo de luz mediante aprendizaje profundo

Emmanuel David Martínez Estrada

Trabajo de Grado para optar al título de Ingeniero de Sistemas

Director

Edwin Mauricio Vargas Díaz

Magister en Ingeniería Eléctronica

Codirector

Henry Arguello Fuentes

Doctor en Ingeniería Eléctrica y Computación

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Bucaramanga

2021

Dedicatoria

A Dios, quien se convirtió en mi fuerza y la causa de todo mi conocimiento, cuyo conocimiento y verdad están por encima de todas las cosas. Esto es solo el principio.

A mi director Edwin Vargas, quien me apoyó durante todo este arduo camino para que este proyecto culminara existosamente, alentándome a no rendirme y a buscar siempre la solución a todos los problemas que se iban presentando.

Al grupo de investigación HDSP, que me brindó los recursos y ayuda necesaria para poder desarrollar este proyecto exitosamente.

A mi familia, por el apoyo incondicional que me han brindado y su infinita paciencia, escuchándome un sinfín de veces mis explicaciones a pesar de que no entender mi trabajo, con el propósito de alentarme a salir adelante.

A mis amigos más cercanos, porque estuvieron apoyando mis decisiones y mis pasos para seguir adelante.

Agradecimientos

A Dios, por darme el deseo y las capacidades para enfrentar todos los desafíos presentes en esta excelente carrera universitaria.

A la Universidad Industrial de Santander, por ser la institución de educación superior que me abrió las puertas a un mejor futuro a través de la educación brindada.

A mi director de tesis, por haberme dedicado tiempo de calidad y haberme guiado con paciencia y esmero a través del desarrollo de este trabajo de grado.

Al grupo de investigación HDSP por brindarme los recursos y herramientas necesarias para desarrollar mi proyecto.

A mi familia, por brindarme incondicional apoyo y motivación para seguir adelante.

Para todos aquellos que creyeron en mi trabajo, a pesar de las dificultades e inconvenientes.

Tabla de Contenidos

Introducción **11**

1 Objetivos **16**

2 Antecedentes Teóricos **17**

2.1 Campos de Luz 17

2.2 Muestreo Compresivo de Imágenes 18

2.3 Aprendizaje Profundo 20

2.3.1 Redes Neuronales Convolucionales (RNCs) 21

2.3.2 Muestreo Descendente y Ascendente 23

2.3.3 Redes Amplias y Profundas 24

2.4 Mapas de Profundidad y Disparidad 26

2.4.1 Estimación de la Profundidad 27

2.4.2 Estimación de la Profundidad basada en Campos de Luz 31

3 Arquitectura Óptica del Campo de Luz **34**

3.1 Adquisición Compresiva de Campos de Luz 34

3.2 Reconstrucción del Campo de Luz a partir de las Medidas Comprimidas 37

4 Estimación de la Profundidad del Campo de Luz Comprimido **41**

4.1 Estimación de la profundidad de Extremo a Extremo 41

4.1.1 Étapa de entrenamiento 42

4.1.2 Étapa de Inferencia 43

4.2 Decodificador Profundo 44

5 Simulaciones y Resultados 49

5.1 Conjunto de Datos 49

5.2 Función de Pérdida y Métricas 51

5.3 Experimentos 52

5.3.1 Configuraciones del Método propuesto 52

5.3.2 Métodos de Comparación 53

5.3.3 Resultados 58

6 Conclusiones 64

Referencias Bibliográficas 65

Lista de Figuras

Figura 1 Representaciones del campo de luz. 17

Figura 2 Representación visual del MC y su reconstrucción. 20

Figura 3 Arquitectura de la U-net original para la segmentación de imágenes biomédicas. 22

Figura 4 Muestreo descendente y ascendente. 23

Figura 5 Tres tipos de arquitecturas para redes neuronales basadas en la distribución de las neuronas o capas convolucionales. 24

Figura 6 Geometría epipolar. 26

Figura 7 Vista central de campos de luz con sus respectivos mapas de disparidad. 28

Figura 8 Diseño del sistema de adquisición de campos de luz con una MCA a una distancia d_m del sensor. 35

Figura 9 Método propuesto con enfoque de extremo a extremo. 41

Figura 10 Modelo para el decodificador del método propuesto. 45

Figura 11 Convolución ascendente rápida 46

Figura 12 Muestras usadas en las configuraciones. 50

Figura 13 Reconstrucciones de campos de luz. 56

Figura 14 Muestras de campos de luz completas y comprimidas. 61

Figura 15 Estimación de los mapas de disparidad. 62

Figura 16 MCAs de tamaño digital real utilizadas en el método propuesto. 63

Lista de Tablas

Tabla 1 Resultados cuantitativos de las reconstrucciones de los campos de luz mediante MC y aprendizaje profundo junto con sus tiempos de entrenamiento e inferencia. 55

Tabla 2 Resultados cuantitativos de la estimación de los mapas de disparidad. 58

Resumen

Título: Estimación de la profundidad a partir de proyecciones 2D del campo de luz mediante aprendizaje profundo[1]

Autor: Emmanuel David Martínez Estrada[2]

Palabras Clave: Estimación de profundidad, campo de luz, muestreo compresivo, redes neuronales convolucionales, enfoque de extremo a extremo.

Descripción: En los últimos años, estimar la profundidad de una escena se ha convertido en una tarea desafiante, debido a que esta información se desvanece al adquirir una única proyección con un sensor bidimensional, generando un problema inverso mal planteado. La profundidad se puede estimar de manera robusta aprovechando la información espacial y angular que proporcionan los campos de luz. Sin embargo, adquirir los campos de luz requieren un alto costo de almacenamiento y de procesamiento limitando el uso de esta tecnología en aplicaciones prácticas. Para superar esta limitación, la teoría de muestreo compresivo ha permitido el desarrollo de arquitecturas ópticas para adquirir una única proyección codificada del campo de luz. Sin embargo, este tipo de técnicas requieren un alto costo computacional para decodificarla. Este trabajo propone optimizar conjuntamente una arquitectura óptica para adquirir el campo de luz a partir de una única proyección y una red neuronal convolucional que funciona como decodificador en un enfoque de extremo a extremo para la estimación de la profundidad. Esto permite estimar directamente la profundidad desde las medidas comprimidas omitiendo el proceso de reconstrucción del campo de luz que se requiere en enfoques tradicionales. Para el decodificador se propone una red compuesta de bloques residuales y proyecciones ascendentes basada en la arquitectura U-net, que contribuye a la estimación óptima de la profundidad a partir de la escasa información que brindan las medidas comprimidas. Experimentalmente, se encontró que el método propuesto estima mapas de disparidad comparables con los obtenidos usando campos de luz reconstruidos. Además, el método propuesto es 20 veces más rápido en el entrenamiento y 23 veces más rápido en la inferencia en comparación con el mejor método que estima la profundidad a partir de campos de luz reconstruidos.

---

[1] Trabajo de Grado

[2] Facultad de Ingeniería Fisicomecánicas. Escuela de Ingeniería de Sistemas. Director: Edwin Mauricio Vargas Díaz. Magister en Ingeniería Eléctronica. Codirector: Henry Arguello Fuentes. Doctor en Ingeniería Eléctrica y Computación.

Abstract

Title: Depth estimation from 2D projections of the light field using deep learning[3]

Author: Emmanuel David Martínez Estrada[4]

Keywords: Depth Estimation, Light Field, Compressive Sensing, Convolutional Neural Networks, end-to-end approach.

Description: In the last years, estimating the depth of a scene has become a challenging task, because this information vanishes when acquiring a single projection with a two-dimensional sensor, generating an ill-posed inverse problem. Depth can be robustly estimated leveraging spatial and angular information provided by light fields. However, acquiring the light fields requires a high cost of storage and processing, limiting the use of this technology in practical applications. To overcome this limitation, the theory of compressive sensing has allowed the development of optical architectures to acquire a single encoded projection of the light field. However, this type of technique requires a high computational cost to decode. This work proposes to jointly optimize an optical architecture to acquire a single light field projection and a convolutional neural network that functions as a decoder in an end-to-end approach for depth estimation. This allows depth to be directly estimated from compressed measurements by omitting the light field reconstruction process required in traditional approaches. For the decoder, a network composed of residual blocks and ascending projections based on the U-net architecture is proposed, which contributes to the optimal estimation of the depth from the little information provided by the compressed measurements. Experimentally, it was found that the proposed method estimates disparity maps comparable with those obtained using reconstructed light fields. Furthermore, the proposed method is 20 times faster in training and 23 times faster in inference compared to the best method that estimates depth from reconstructed light fields.

---

[3] Degree work

[4] Faculty of Physicomechanical Engineering. School of Systems Engineering. Director: Edwin Mauricio Vargas Díaz. Master in Electrical Engineering. Codirector: Henry Arguello Fuentes. Doctor in Electrical and Computer Engineering.

Introducción

Los campos de luz recolectan la cantidad de luz proveniente de todas las direcciones en cada punto espacial de una escena física, siendo diferente a la fotografía tradicional, en la cual se adquieren proyecciones 2D de la luz de una escena tridimensional. Estos campos de luz se pueden representar por medio de la función plenóptica Adelson et al. (1991), o mediante la parametrización de dos planos Levoy and Hanrahan (1996); Gortler et al. (1996). El análisis de los campos de luz ha permitido el desarrollo de diversas aplicaciones como reducir la oclusión, modificar el desenfoque de imágenes, o la creación de imágenes 3D. Sin embargo, adquirir estos campos de luz implica grandes costos de almacenamiento debido al gran volumen de infomación espacial y angular. Adicionalmente, la multiplexación en micro lentes de algunos de los sistemas ópticos también impone una compensación entre la resolución espacial y angular Wu et al. (2017).

Para superar estas limitaciones, se han desarrollado diferentes arquitecturas ópticas para la adquisición de muestras comprimidas de campos de luz basada en la teoría del muestreo compresivo Marwah et al. (2013); Inagaki et al. (2018); Hajisharif et al. (2020). El muestreo compresivo de imágenes aprovecha el hecho de que si los campos de luz son vistos como un conjunto de imágenes de sub-apertura, es decir, las imágenes captadas por una cámara, estos son escasos en alguna base de representación y por lo tanto, pueden ser comprimidos. Esto representa una gran ventaja en el tratamiento de los datos en aplicaciones prácticas porque al realizarse estas mediciones se reducen los costos de almacenamiento. Tradicionalmente, estas muestras comprimidas pueden ser

reconstruidas mediante diversos algoritmos que son computacionalmente costosos. Por lo tanto, si se omite este proceso de reconstrucción mediante algoritmos y se trabaja directamente sobre las medidas comprimidas, se reduce la propagación de errores y además mejora la velocidad de procesamiento de algoritmos.

Por otra parte, estimar la profundidad se ha convertido en una tarea computacional que permite simular entornos tridimensionales virtuales más realistas que las imágenes convencionales 2D o imágenes panorámicas. Algunas de sus aplicaciones más relevantes son remodelado computacional 3D de escenarios de la vida real El Gendy et al. (2011); Huang et al. (2019), el reconocimiento de rostros Chen and Chellappa (2017), la robótica Sawano et al. (2001); De Cubber and Doroftei (2011) y la medicina Nam et al. (2012); Wu et al. (2020). La representación tradicional de los mapas de profundidad viene dada por la geometría epipolar, donde tanto los mapas de disparidad como los de profundidad pueden ser calculados a partir de la relación geométrica existente entre las diferentes imágenes de sub-apertura adquiridas de los campos de luz. Existen múltiples propuestas para resolver esta tarea que se basan en imágenes monoculares Godard et al. (2017, 2019); Garg et al. (2019), estereoscópicas estáticas Smolyanskiy et al. (2018); You et al. (2019); Shen et al. (2021) o en movimiento Zhou et al. (2017); Mahjourian et al. (2018); Casser et al. (2019).

Una forma de mejorar la estimación de la profundidad es aprovechando la información angular proporcionada por las múltiples vistas de los campos de luz mediante diferentes métodos. Por un lado, existen métodos basados en la correspondencia de imágenes de sub-apertura, que pre-

sentan dificultades debido a la estrecha distancia de disparidad presente en las imágenes, esto se resuelve a partir de la explotación diversas señales presentes en los campos de luz Lin et al. (2015); Zhu et al. (2017). Por otro lado, están los métodos basados en la geometría epipolar que aprovechan las propiedades de la estructura de los campos de luz para realizar una óptima estimación Sheng et al. (2018); Li and Jin (2020). Sin embargo, aún existen técnicas basadas en el aprendizaje profundo que no han sido exploradas para realizar esta tarea.

En los últimos años, el aprendizaje profundo ha tomado gran relevancia en el campo de la investigación permitiendo obtener resultados superiores a métodos tradicionales Zhang et al. (2018). Una de las ramas más importantes del aprendizaje profundo es el aprendizaje supervisado, especialmente, los métodos basados en las redes neuronales convolucionales (RNCs) Albawi et al. (2017). La mayor ventaja que ofrecen las RNCs subyace en su capacidad para extraer características correlacionadas de los datos y combinarlas para permitir un aprendizaje más rápido, eficiente y automático, lo cual reduce considerablemente el tiempo de procesamiento de datos y mejora la calidad de los resultados. Debido a esto, las RNCs también se pueden aprovechar para estimar la profundidad a partir de métodos tradicionales Sanz et al. (2012); Godard et al. (2017, 2019); Smolyanskiy et al. (2018); Shen et al. (2021); Zhou et al. (2017); Mahjourian et al. (2018) o en el uso de campos de luz Wu et al. (2017); OMahony et al. (2019); Mun and Ho (2018); Li et al. (2020b, 2021a).

Una metodología empleada recientemente para resolver diversas tareas computacionales

consiste en el diseño de componentes ópticos mediante las metodologías del aprendizaje profundo. Específicamente, existen varias técnicas basadas en el aprendizaje profundo para el diseño de aperturas codificadas (ACs) o máscaras codificadas (MCOs), donde sus patrones son aprendidos para diferentes campos de aplicación según la información de la luz requerida. Dentro de los cuales se destacan la información espectral, temporal, espacial, polarización, profundidad, amplitud, fase y vistas angulares Shedligeri et al. (2017); Haim et al. (2018); Chang and Wetzstein (2019); Wu et al. (2019); Wang et al. (2018); Bacca et al. (2020, 2021).

En el estado del arte se han desarrollado métodos para estimar la profundidad mediante la metodología mencionada que solo se basan en imágenes 2D codificadas con diferentes diseños de AC utilizando un kernel de desenfoque y de fase óptica Shedligeri et al. (2017); Haim et al. (2018). Adicionalmente, existen métodos que resuelven esta tarea a través del aprendizaje profundo mediante la reconstrucción de medidas comprimidas Liu et al. (2018, 2020) ó, inclusive utilizando enfoque de extremo a extremo (EAE) donde se implementan arquitecturas ópticas para la adquisición de imágenes monoculares Shedligeri et al. (2017); Haim et al. (2018); Chang and Wetzstein (2019); Wu et al. (2019). Los trabajos que tienen mayor relación con el método propuesto estiman la profundidad como un proceso intermedio para reconstruir campos de luz a partir de medidas comprimidas no entrenables Vadathya et al. (2017, 2019). Desde el conocimiento de los autores de este trabajo, no se ha reportado en el estado del arte un método que estime la profundidad directamente de las medidas comprimidas de los campos de luz.

En este trabajo se propone un algoritmo para estimar la profundidad de los campos de luz

comprimidos basado en RNCs mediante un enfoque EAE que no requiere la reconstrucción de las escenas. En particular, la metodología propuesta se compone de dos etapas: la etapa de entrenamiento y la etapa de inferencia. En la etapa de entrenamiento, se divide la red neuronal en dos capas: la primera capa es representada como un modelo diferenciable que modela el sistema de fotografía propuesto por Marwah et al. (2013), que consiste en adquirir las escenas comprimidas y la segunda capa está representada por un decodificador basado en redes neuronales profundas, que se encarga de estimar la profundidad de las escenas codificadas adquiridas.

1. Objetivos

Objetivo general

Desarrollar un algoritmo que permita estimar los mapas de profundidad a partir de proyecciones 2D comprimidas de alta resolución espacial y angular del campo de luz mediante el uso de redes neuronales profundas.

Objetivos específicos

Describir matemáticamente la función de transferencia del sistema de adquisición compresiva del campo de luz;

Implementar un algoritmo computacional para la simulación de la adquisición del sistema de medidas comprimidas del campo de luz;

Implementar un algoritmo basado en RNC para estimar mapas de profundidad a partir de medidas comprimidas del campo de luz basado en aprendizaje profundo;

Evaluar el rendimiento del algoritmo propuesto para la estimación de mapas de profundidad a través de diferentes métricas de error y calidad de resultados.

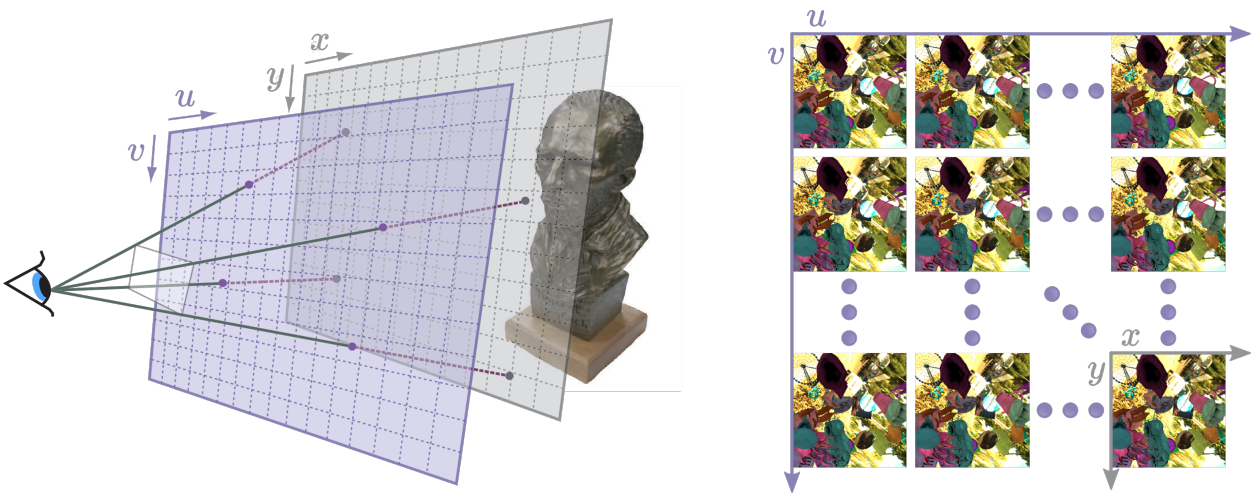
2. Antecedentes Teóricos

2.1. Campos de Luz

Los seres humanos tienen la capacidad de visualizar el mundo que les rodea gracias a la luz que incide en los diferentes objetos de la naturaleza, y debido a que cuentan con un par de ojos pueden obtener la información tridimensional del entorno. Particularmente, estos rayos de luz son conocidos como campos de luz, los cuales contienen información espacial y angular. Los campos de luz pueden ser aprovechados para realizar múltiples tareas como reducir la oclusión, crear imágenes 3D, segmentar diferentes objetos, reconocer objetos o rostros, detección, clasificación, estimación de profundidad, entre otros Wu et al. (2017). En 1991, Bergen dio una definición formal para representar la luz mediante la función plenóptica Adelson et al. (1991), esta función

Figura 1

Representaciones del campo de luz.



Nota: De izquierda a derecha: parametrización de dos planos Chen (2003), y matriz de imagenes de sub apertura (SAI) Schambach and Heizmann (2020).

$L(x,y,z,\theta,\phi,\lambda,t)$ describe un entorno multidimensional a través de las múltiples perspectivas de la escena bajo estudio, donde (x,y,z) representan las coordenadas espaciales de cada rayo de luz incidente en la escena, (θ,ϕ) representa cada posible ángulo en la escena en coordenadas esféricas, λ representa el espectro electromagnético para cada una de las coordenadas espaciales dadas y finalmente, t representa el tiempo. El campo de luz surge de la función plenóptica al considerar únicamente las coordenadas espacial y angular de la luz como $L(x,y,z,\theta,\phi)$, estudios posteriores lograron reducir esta representación a 4D Levoy and Hanrahan (1996); Gortler et al. (1996). Esta nueva representación de la función plenóptica viene dada por dos planos donde los rayos inciden de un plano al otro y se modela matemáticamente como $L(x,y,u,v)$ donde el primer plano coordenado está determinado por (x,y), el segundo plano por (u,v) y los rayos de luz en este sistema inciden del plano *uv* al plano *xy*, este modelo también es conocido como la parametrización de dos planos. Otra forma de representar los campos de luz consiste en interpretarlos como matrices de múltiples vistas, llamadas imágenes de sub apertura (SAI), como se muestra en la figura 1.

2.2. Muestreo Compresivo de Imágenes

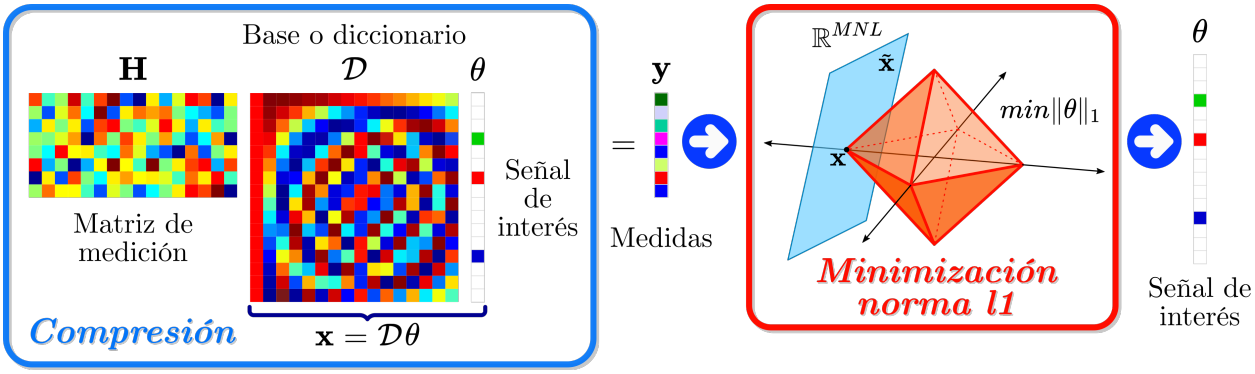
El muestreo compresivo (MC) surgió como una novedosa técnica de adquisición y reconstrucción de datos en condiciones de escasez, permitiendo una compresión directa y conduciendo a una implementación de hardware eficiente. Debido a sus capacidades, el MC ha sido ampliamente aplicado en la compresión de datos, el cifrado de imágenes, criptografía, reconstrucción de redes complejas, estimación de canales, conversión de analógico a digital, codificación de canales, reconstrucción de radar, entre muchas otras aplicaciones Li et al. (2020a). El MC confía en dos principios fundamentales: la baja densidad y la incoherencia Candès and Wakin (2008). La baja

densidad se refiere a que la señal de interés puede ser representada por muy poca información y la teoría sobre el MC de imágenes explotan este hecho en el sentido de que las imágenes convencionales pueden ser de baja densidad cuando son representadas en una base apropiada, también conocida como diccionario $\mathcal{D}$. Matemáticamente, si representamos una imagen como un vector $\mathbf{f} \in \mathbb{R}^{MNL}$, donde $\mathbf{f}$ es la representación vectorial de un tensor $\mathbf{F} \in \mathbb{R}^{M \times N \times L}$, de tal manera que tiene una resolución espacial de $M \times N$ y una profundidad espectral L, entonces, esto puede expresarse como $\mathbf{f} = \mathcal{D}\theta$, donde θ es una representación vectorial de baja densidad con k elementos distintos de cero y $k \ll MNL$. Por otra parte, la incoherencia expresa la idea de que los vectores de baja densidad en $\mathcal{D}$ también deben ser de baja densidad en el dominio que son adquiridos Candès and Wakin (2008). El modelo tradicional del MC se puede representar matricialmente como $\mathbf{y} = \mathbf{Hf}$, $\mathbf{y} \in \mathbb{R}^k$, y donde $\mathbf{H}$ es una representación matricial del sistema de muestreo de la arquitectura óptica para la adquisición de imágenes y la representación final de la adquisición de medidas es $\mathbf{y} = \mathbf{H}\mathcal{D}\theta$. Como se observa en la figura 2, la forma tradicional de reconstruir la imagen original está representada por θ que se obtiene al resolver $\tilde{\mathbf{f}} = \mathcal{D}\left(\underset{\theta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{H}\mathcal{D}\theta\|_2^2 + \lambda\|\theta\|_1\right)$.

El uso del MC también se ha extendido a los campos de luz debido a su gran variedad de aplicaciones Qaisar et al. (2013). Por ejemplo, Kshitij et al. Marwah et al. (2013) presentaron un sistema de reconstrucción de baja densidad de campos luz comprimidos combinando proyecciones ópticamente codificadas y reconstrucciones computacionales no lineales bajo el uso de diccionarios optimizados llamados átomos del campo de luz. Otro método propuesto consiste en adquirir los campos de luz mediante cámaras de apertura codificada (AC) o máscara codificada (MCO) haciendo uso del aprendizaje profundo Inagaki et al. (2018). También se ha podido extender la

Figura 2

Representación visual del MC y su reconstrucción.



aplicación del MC de imágenes al dominio temporal, permitiendo la captura de videos de campos de luz usando un solo sensor Hajisharif et al. (2020).

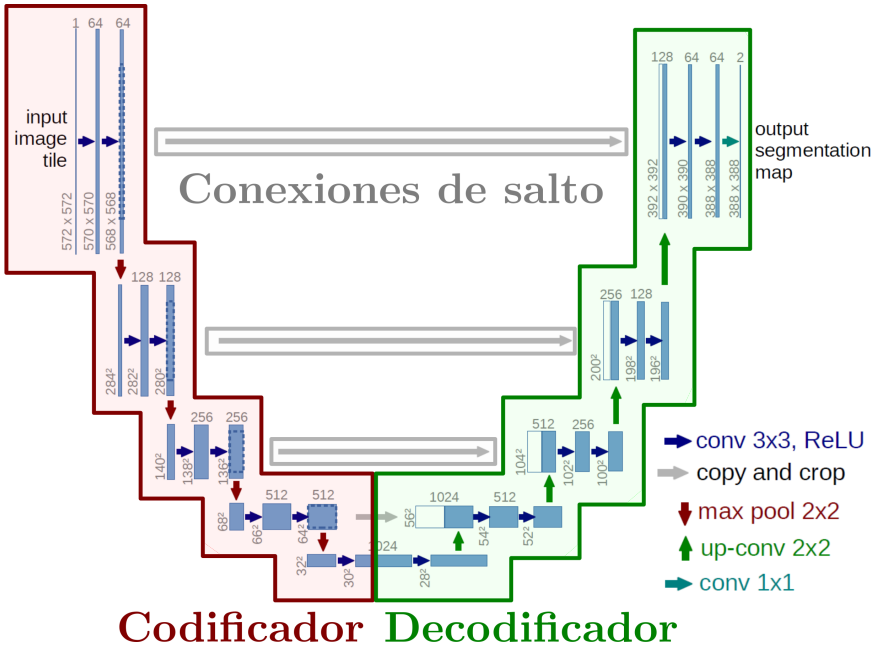
2.3. Aprendizaje Profundo

El aprendizaje profundo surgió como una poderosa herramienta para resolver diversas tareas basadas en algoritmos computacionales que aprovechan la gran cantidad de datos digitales en el mundo que han generado en los últimos años Zhang et al. (2018). El objetivo de estos algoritmos computacionales consiste en aprender características de los datos de forma progresiva para deducir nueva información, los resultados obtenidos con estos métodos en los diferentes tipos de tareas computacionales han demostrado ser superiores a los métodos tradicionales OMahony et al. (2019). El aprendizaje profundo consta de diversas aplicaciones tales como la conducción autónoma, imágenes médicas, reconocimiento automático de imágenes, entre otros Balas et al. (2019). El aprendizaje profundo se puede dividir en dos ramas de acuerdo con la disposición del conjunto de entrenamiento: el aprendizaje supervisado y el aprendizaje no supervisado. En el aprendiza-

je supervisado, el algoritmo de aprendizaje puede ser visto como una función que recibe datos etiquetados como entrada y aprende a generar una salida adecuada a partir de las etiquetas, esto permite generalizar salidas más allá de los datos etiquetados conocidos, la estrategia más utilizada para este tipo de aprendizaje son las RNCs Albawi et al. (2017). Formalmente, dado un conjunto de entrenamiento con M muestras $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{M}$, donde $\mathbf{X}_i$ representa la imagen de entrada y $\mathbf{Y}_i$ la respectiva etiqueta o salida, esto también puede ser visto como una matriz de salida dependiendo del tipo de tarea a resolver, por ejemplo, como clasificación o segmentación. El objetivo consiste en aprender un mapeo no lineal $\mathcal{N}_\theta$, donde θ representa los parámetros entrenables de la función. Esto se resuelve mediante algoritmos de optimización, una solución sencilla sería resolver $\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(\mathbf{Y}_i, f_\theta(\mathbf{X}_i))$, donde $\mathcal{L}$ es una función de costo o pérdida que se ajusta al mapeo no lineal f_θ y los pesos óptimos del modelo son representados por θ^*. Por otra parte, en el aprendizaje no supervisado los conjuntos de entrenamiento no poseen etiquetas $\mathbf{Y}_i$. Por lo tanto, a partir de $\mathbf{X}_i$, los algoritmos aprenden sobre los datos interpretando su distribución y realizan inferencias automáticas. Por ejemplo, el algoritmo k-means aprende a clasificar los datos mediante el uso de k centroides MacQueen et al. (1967). En relación con el método propuesto, en las siguientes subsecciones se mostrará la relevancia de las RNCs en tareas computacionales relacionadas con el procesamiento de imágenes y algunas de las arquitecturas de redes neuronales más relevantes del estado del arte relacionados con la capa profunda propuesta. Se mostrarán algunos métodos para realizar muestreo descendente y ascendente de los cuales dependen las RNCs para aprender una representación latente y realizar la decodificación de los datos. Finalmente, se mostrarán que tipos de redes se pueden diseñar de acuerdo con su cantidad de capas convolucionales y su distribución.

Figura 3

Arquitectura de la U-net original para la segmentación de imágenes biomédicas.



Nota: Propuesta por Ronneberger et al. (2015).

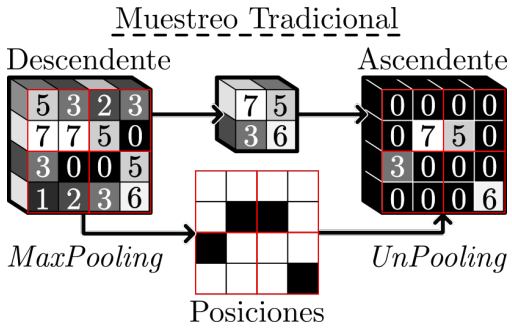
2.3.1. Redes Neuronales Convolucionales (RNCs). Las RNCs tienen la capacidad de detectar las características más relevantes de los conjuntos de datos de entrenamiento aprovechando la coherencia espacial, debido a que, todas las capas comparten la información que van adquiriendo sin supervisión humana. Las RNCs son usadas para resolver tareas de aprendizaje profundo relacionadas con la clasificación, segmentación, detección de objetos, reconocimiento de rostros, entre otros Li et al. (2021b). La red U-net es ampliamente conocida por su estructura de codificación donde las características de las imágenes son aprendidas hasta llegar a resolución espacial bastante reducida, donde se genera una representación codificada de la imagen para una posterior decodificación mediante muestreo ascendente. Por último, las capas de salto permite

conservar la consistencia espacial de decodificación al incluir la codificación de acuerdo a la resolución espacial de las capas, como se observa en la figura 3. U-net ha sido ampliamente estudiada en la tarea de segmentación con múltiples variaciones propuestas, como lo son Ni et al. (2019); Chen et al. (2019); Huang et al. (2020) por mencionar algunos, pero este rendimiento también se ha extendido a los otros tipos de tareas computacionales como lo son Guan et al. (2019); Valloli and Mehta (2019); Qin et al. (2020). En este trabajo se propone una RNC que también está basada en U-net, inspirada en Laina et al. (2016); Harsányi et al. (2018), los detalles del modelo serán explorados en la sección 4.2.

2.3.2. Muestreo Descendente y Ascendente. Muchas redes neuronales requieren de un sistema de reducción y ampliación espacial del mapeo de características. Como se observa en la figura 4, el objetivo del muestreo descendente consiste en reducir espacialmente el mapa de características a una dimensión definida por el tamaño del muestreo. Por ejemplo, cuando el muestreo descendente es 2×2, entonces se toman regiones espaciales con esa dimensión y se extrae un

Figura 4

Muestreo descendente y ascendente.

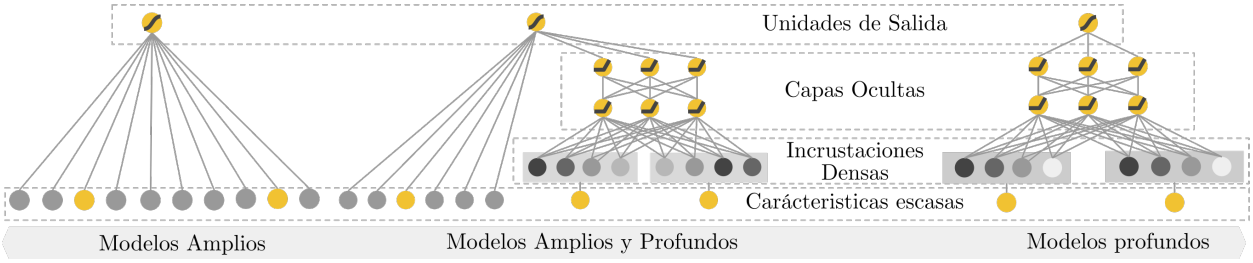


Nota: Ejemplo de muestreo descendente 2×2 *basado en MaxPooling de un mapa de carácteristicas 2D de dimensiones* 4×4 *y muestreo ascendente del nuevo mapa de carácteristicas basado en UpPooling* 2×2.

solo valor espacial. Existen varias formas de realizar esta tarea, como calcular el promedio de cada región, extraer el mínimo valor o extraer el máximo valor Li et al. (2021b). En este caso, se aplica la operación *MaxPooling* que extrae el mayor valor de cada región y almacena la posición, esto permite detectar las regiones de cambio de color en el mapa de características, como lo son bordes. Por otro lado, el objetivo del muestreo ascendente *UnPooling* consiste en expandir espacialmente el mapa de características dado un tamaño de muestreo, como se observa en la figura 4. Esta tarea se realiza reacomodando los valores del mapa de características en las posiciones almacenadas por el muestreo descendente. Sin embargo, como estos nuevos mapas de carácteristicas suelen ser operados con diversas capas convolucionales, entonces podría ser contraproducente al aprendizaje de una red neuronal. Para evitarlo, Laina et al. Laina et al. (2016), propusieron un muestreo ascendente que divide este muestreo en operaciones convolucionales, cuyos resultados son entrelazados para formar el mapa de características final.

Figura 5

Tres tipos de arquitecturas para redes neuronales basadas en la distribución de las neuronas o capas convolucionales.



Nota: Tomado de Cheng et al. (2016).

2.3.3. Redes Amplias y Profundas. Generalmente, se intentan desarrollar redes neuronales que sean profundas, es decir, que tengan cientos de capas convolucionales, debido a que se presume que esto debería mejorar los resultados, no obstante, se ha demostrado que no siempre se cumple este hecho He et al. (2016). Cuando las redes neuronales son demasiado profundas se presenta el problema del desvanecimiento del gradiente, donde los pesos se vuelven valores tan pequeños que son imposibles de optimizar correctamente. En el peor caso, los parámetros de las redes neuronales no convergerán. La solución que fue propuesta por He et al. (2016) define los bloques residuales, cuya intuición principal se centra en el mapeo residual. Formalmente, se desea alcanzar un mapeo de características $\mathcal{H}(x)$, siendo x el mapa de características actual en la red. Si $\mathcal{F}(x) = \mathcal{H} - x$ es el residuo entre la entrada y la salida de estas capas, entonces el mapeo deseado sería $\mathcal{H} = \mathcal{F} + x$, esto también es conocido como un mapeo de identidad. Otras variaciones de estos bloques hacen uso de capas convolucionales en la entrada x para lograr un mejor rendimiento de las redes. Sin embargo, la solución de He et al. (2016) propone una red neuronal con bloques residuales (ResNet) que sigue siendo extremadamente profunda. Por lo tanto, Zagoruyko and Komodakis (2016) diseña una variante de ResNet basándose en reducir la profundidad y ampliar la red, es decir, añadir más filtros en las convoluciones. Este método ha sido utilizado para resolver diversas tareas de aprendizaje profundo como segmentación semántica Nakayama et al. (2018), detección de objetos en tiempo real Lee et al. (2017), super resolución Deeba et al. (2021), entre otros. Otra solución puede presentarse al aprovechar conjuntamente las bondades de las redes profundas y amplias, como lo sugiere Cheng et al. (2016), donde se combina la memorización y generalización

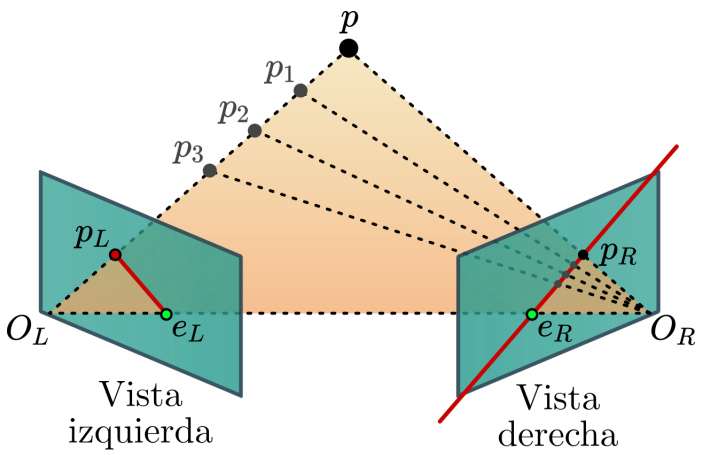
de los sistemas que recomiendan. En la figura 5 se puede observar los esquemas de los 3 tipos de redes neuronales.

2.4. Mapas de Profundidad y Disparidad

Los mapas de profundidad o disparidad surgen de la visión en estéreo por computador donde se busca extraer información del entorno 3D de imágenes digitales. Esto se logra al colocar dos cámaras a una relativa distancia apuntando hacia la misma escena, simulando los ojos humanos. Al realizar esta configuración, se consigue una serie de relaciones geométricas entre los puntos 3D de la escena y las proyecciones 2D de las imágenes captadas por ambas cámaras.

Figura 6

Geometría epipolar.



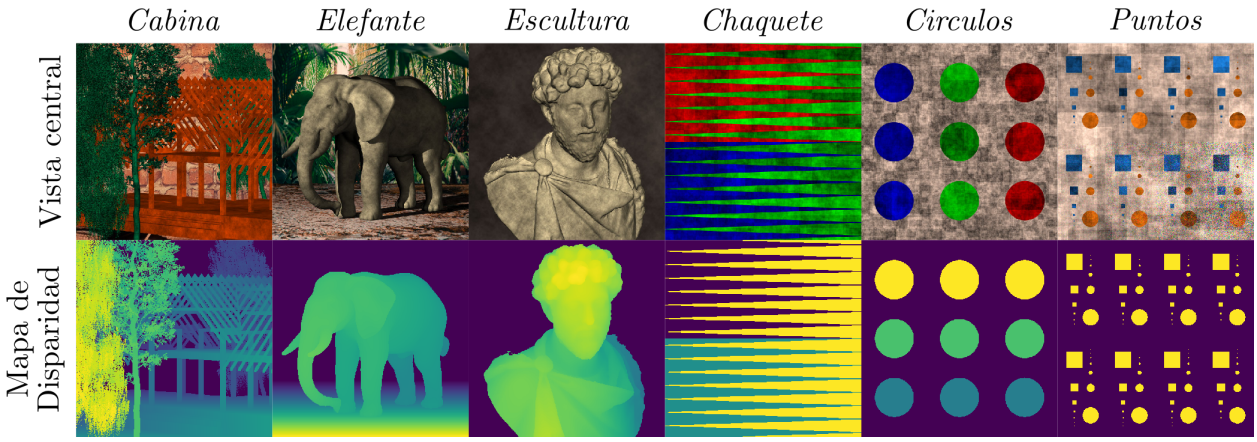
Formalmente, este sistema puede ser descrito a través de la geometría epipolar Zhang (1998) como se observa en la figura 6. Si se tienen dos cámaras observando el mismo punto 3D **p** y el centro de simetría de ambas cámaras vienen dadas por $\mathbf{O_L}$ y $\mathbf{O_R}$, respectivamente. Entonces la proyección de dicho punto **p** sobre los planos imagen de ambas cámaras son $\mathbf{p_L}$ y $\mathbf{p_R}$. Debido a que los centros ópticos de las cámaras son distintos, entonces se obtiene una proyección distinta

de la misma escena en cada plano imagen. Como se observa en la figura 6, $\mathbf{p_L}$ es la proyección de los puntos $\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{p_3}$, ... en el plano imagen de la izquierda. Por otra parte, estos mismos puntos $\mathbf{p_i}$ se proyectan en el plano imagen de la derecha formando una línea conocida como **línea epipolar**. La línea $\mathbf{O_L O_R}$ es conocida como la **base**. El plano definido por $\mathbf{O_L p O_R}$ es conocido como **plano epipolar** y finalmente, los puntos donde la base intercepta con los planos imagen $\mathbf{e_L}$ y $\mathbf{e_R}$ son conocidos como **epipolos**. La disparidad se calcula al estimar la diferencia en la posición en el plano imagen del mismo punto 3D **p**, es decir, calcular la diferencia entre p_L y p_R, y el mapa de disparidad se obtiene cuando este proceso se realiza a cada uno de los puntos proyectados en los planos imagen de la manera $D(x,y) = p_L(x,y) - p_R(x,y)$, siendo $D(x,y)$ la disparidad en la coordenada (x,y) de cada plano imagen (ver figura 7). Por otra parte, los mapas de profundidad se obtienen mediante $d(x,y) = bf/D(x,y)$, donde b es la base y f es la distancia focal de las cámaras Jain et al. (1995). En las siguientes subsecciones se mencionarán los métodos más comunes para estimar la profundidad, desde imágenes monoculares hasta visión en estéreo, video y campos de luz.

2.4.1. Estimación de la Profundidad. Conocer la profundidad de un escenario tridimensional tiene un papel fundamenta en diversas aplicaciones en campos relacionados con el procesamiento de imágenes, visión por computador, modelado 3D, robótica, realidad virtual, aumentada y mixta, entre otros campos El Gendy et al. (2011); Huang et al. (2019); Nam et al. (2012); Wu et al. (2020); Kim and Sohn (2003); Kalia et al. (2019); El Jamiy and Marsh (2019); Xu et al. (2020); Diamantas et al. (2010); Nalpantidis and Gasteratos (2012); Ye et al. (2017). Los principales métodos para la estimación de profundidad se enfocan en estimar ya sea los mapas

Figura 7

Vista central de campos de luz con sus respectivos mapas de disparidad.



Nota: Tomado de Schambach and Heizmann (2020).

de disparidad o los mapas de profundidad, dado que tienen una relación geométrica que permite calcular un tipo de mapa a partir del otro. Sin embargo, la complejidad de esta tarea puede aumentar debido a la textura de los elementos en las imágenes, la oclusión o resolución de soluciones ambiguas, donde se podría obtener la misma imagen de diferentes planos imagen. Por tal razón, se han planteado múltiples métodos para estimar la profundidad, donde la mayoría están basadas en el aprendizaje profundo. A continuación se describen algunos de estos métodos:

- **Visión en estéreo:** Es la forma tradicional de estimar la profundidad. Computacionalmente, se usan dos cámaras desplazadas una corta distancia entre ellas que observan la misma escena, una vez adquirida la escena se aprovechan las relaciones geométricas proporcionadas por las imágenes y el entorno físico para extraer la información tridimensional de la escena original. Por mencionar algunos, Wang et al. Wang et al. (2019) proponen la arquitectura AnyNet (Anytime Stereo Network) centrado en estimar mapas de disparidad en tiempo real

en dispositivos móviles, esta arquitectura está basada en U-net como extractor de características y redes de disparidad para cada uno de los niveles de extractor sea adicionado para obtener la estimación final. Por otro lado, Tankovich et al. Tankovich et al. (2021) proponen una red neuronal enfocada en la coincidencia estéreo que se enfoca principalmente en una rápida reconstrucción inicial, seguida de una propagación geométrica bidimensional y mecanismos de deformación para inferir hipótesis de disparidad.

- **Imágenes monoculares:** En esta metodología una sola imagen es usada para estimar la profundidad. Generalmente, se requiere de RNCs preentrenadas para obtener óptimos resultados al momento de realizar esta tarea. Laina et al. Laina et al. (2016) aprovecha el entrenamiento previo de algunas RNCs del estado del arte Sapijaszko and Mikhael (2018) y proponen la capa de proyecciones ascendentes para optimizar la decodificación de los datos en una RNC completamente conectada. Por otro, Harsanyi et al. Harsányi et al. (2018) realizan un caso de estudio donde agregan conexiones de salto para formar una red U-net con bloques residuales y las proyecciones ascendentes. La red propuesta en este trabajo está inspirada en este último trabajo, a excepción de la estructura de la red, la cual será menos profunda, más amplia y simétrica. Además, la RCN propuesta no depende de ninguna RNC preentrenada.

- **Video:** En este caso, la estimación de la profundidad adquiere una nueva dimensión que se encuentra en el dominio temporal, donde se estima la profundidad a una cantidad finita de imágenes correlacionadas. Por mencionar algunos trabajos, Mahjourian et al. Mahjourian et al. (2018) proponen una RNC sin supervisión para estimar la profundidad de las escenas

y un robot de *egomotion* a partir de videos monoculares, donde se promueve la inferencia de la geometría tridimensional de la escena entera y se refuerza la consistencia al estimar una nube de puntos tridimensional a través de la secuencia finita de imágenes en las escenas. Adicionalmente, Casser et al. Casser et al. (2019) proponen una mejora a la RNC mencionada con anterioridad al incluir estimadores de movimiento de los objetos en las escenas y añadir un modelo de refinamiento en línea.

Con mayor relación al método propuesto en este trabajo, algunos trabajos desarrollados se encuentran enfocados en la estimación de la profundidad a partir del dominio comprimido de datos mediante el diseño de ACs con un enfoque EAE. Por mencionar algunos, Shedligeri et al. Shedligeri et al. (2017) proponen un enfoque basado en el aprendizaje de un patrón óptimo de AC a partir de una sola imagen codificada. Igualmente, Harel et al. Haim et al. (2018), proponen una cámara con apertura de fase codificada, la cual provee características de color inequívocas relacionadas con la profundidad de la imagen capturada. Adicionalmente, también se pueden añadir señales de profundidad mediante un desenfoque borroso codificado para una estimación de profundidad monocular, como lo proponen Chang et al. Chang and Wetzstein (2019). Finalmente, Wu et al. Wu et al. (2019) proponen un marco de trabajo para una optimización con enfoque EAE donde también se aprende una máscara de fase óptima junto con una RNC, donde se implementó un sistema físico llamado PhaseCam3D. Todos los ejemplos anteriores contienen RNCs con estructuras parecidas, donde se usan capas convolucionales y capas completamente conectadas. Estos trabajos solo contienen enfoques basados en estimaciones de profundidad a partir de imágenes monoculares. En la siguiente subsección se verá la estimación de la profundidad basada en campos de luz.

2.4.2. Estimación de la Profundidad basada en Campos de Luz. Como los campos de luz son una representación de mayor dimensionalidad que contienen subconjuntos de imágenes de una sola escena, esto permite que la estimación de la profundidad sea posible. La premisa consiste en estimar mapas de disparidad de las imágenes subyacentes y estimar una profundidad única de los mapas obtenidos.

Algunos investigadores han aprovechado las propiedades ya establecidas de la geometría epipolar que puede ser generada directamente del campo de luz. Sheng et al. Sheng et al. (2018) desarrollaron un método de imágenes epipolares de orientación múltiple y extracción de la profundidad teniendo en cuenta la oclusión. Por otra parte, Junke Li et al. Li and Jin (2020) propusieron un algoritmo para estimar la profundidad basado en la distribución de vecindad de cortes de imágenes epipolares. Otro método llamado coincidencia estéreo consiste en calcular la disparidad para cada píxel en el par de imágenes de referencia, imágenes del campo de Luz. Kang Zhu et al. Zhu et al. (2018) desarrollaron un sistema de adquisición de campo de luz con extracción del mapa de profundidad. La técnica de coincidencia estéreo también ha sido usada para estimar un mapa de profundidad inicial y a través del aprendizaje profundo es refinada en el trabajo desarrollado por Rogge et al. Rogge et al. (2020).

Con el propósito de aprovechar el muestreo compresivo, Xiaomin et al. Liu et al. (2018) desarrollaron un sistema adquisición de campo de luz comprimido y se estimó la profundidad con fusión de pistas múltiples y en Liu et al. (2020) también se desarrolló un sistema similar, pero esta vez la profundidad se estimó basado en la síntesis de colores verdaderos y la fusión de información

múltiple. Sin embargo, la estimación de la profundidad la hace a partir de la reconstrucción de las medidas comprimidas y tampoco se observa el uso de aprendizaje profundo en ninguno de los dos trabajos.

A través de aprendizaje profundo Shin et al. Shin et al. (2018) explotan las propiedades de la geometría epipolar y el aumento de muestras para estimar la profundad. Sin embargo, no estima correctamente regiones reflectivas o metálicas. Ma et al. Ma et al. (2018) resuelven este problema introduciendo convoluciones atroces que amplian el campo receptivo de las RCNs sin aumentar el número de parámetros. Por otro lado, Ji-Hun Mun et al. Mun and Ho (2018) propusieron estimar la profundidad removiendo el problema de la discontinuidad de los métodos tradicionales a través de los bloques residuales de las RNCs. Tomando ventaja de las múltiples vistas de los campos de luz, Tsai et al. Tsai et al. (2020) proponen una red neuronal con módulos de selección de vista que generan mapas de atención para extraer la información más relevante y explotan la propiedad de simetría de las vistas de los campos de luz para lograr una mejor precisión. Por otro lado, surge un método llamado Manet, una red neuronal que estima la profundidad del campo de luz a través de una estructura jerárquica de multiescala, entrenada de extremo a extremo, creada por Yan Li et al. Li et al. (2020b). Recientemente, este mismo autor publicó un método para estimar la profundidad a partir de redes neuronales convolucionales basándose en los campos de luz de línea de base amplia Li et al. (2021a). Sin embargo, no se hace uso del muestreo compresivo.

Por otra parte, el método desarrollado por Vadathya et al. Vadathya et al. (2017) se divide en tres pasos. El primer paso consiste en estimar las vistas centrales completamente en foco de las medidas comprimidas de los campos de luz mediante una CA no entrenable usando una RNC.

El segundo paso consiste en estimar los mapas de disparidad de los campos de luz concatenando las vistas centrales estimadas con las respectivas medidas comprimidas mediante otra RNC. Finalmente, el tercer paso consiste en reconstruir los campos de luz mediante la deformación de las vistas centrales estimadas usando los mapas de disparidad estimados. La segunda RNC usada para la disparidad no requiere de los valores verdaderos, puesto que, se entrena de forma conjunta con el tercer paso. Este método es extendido por el mismo autor al estimar campos de disparidad en el segundo paso Vadathya et al. (2019). Estos son los métodos que tienen mayor relación con el método propuesto en este trabajo. Cabe resaltar que, nuestro trabajo también utiliza un enfoque con redes neuronales que se basa en una estimación directa de la profundidad sin reconstruir ningún campo de luz.

En los siguientes capítulos se mencionará el sistema de adquisición de la arquitectura óptica y como se incorpora directamente a una RNC para formar un modelo entrenable de extremo a extremo. En el caso de la estimación de la profundidad, varios autores Shedligeri et al. (2017); Haim et al. (2018); Chang and Wetzstein (2019); Wu et al. (2019) han demostrado que es plausible resolver esta tarea directamente de las medidas comprimidas con un correcto diseño de las ACs y/o MCOs.

3. Arquitectura Óptica del Campo de Luz

El diseño de arquitecturas ópticas surge como una necesidad de resolver diversos problemas relacionados con tareas computacionales resultando en novedosas soluciones que extraen información adicional de diferentes tipos de escenas, ya sea para inferir sobre nueva información o realizar análisis más complejos. Con el nacimiento del MC surgen múltiples arquitecturas ópticas donde se aprovecha el sensado de información continua en un dominio escaso Li et al. (2020a). Las arquitecturas ópticas de MC hacen uso de AC o MCO para proyectar la información de la escena codificada sobre un sensor, donde es posteriormente recuperada mediante métodos de reconstrucción Pope (2009). Recientemente han surgido arquitecturas ópticas que pueden ser diseñadas mediante redes neuronales con un enfoque EAE, donde los componentes ópticos pueden ser simulados para aprender sus características de acuerdo con la tarea computacional que realicen Bacca et al. (2021). En este capítulo se dará un vistazo a la arquitectura de interés para el método propuesto que busca recolectar medidas comprimidas de campos de luz y se observará tanto el método de reconstrucción tradicional basado en MC como en aprendizaje profundo.

3.1. Adquisición Compresiva de Campos de Luz

La adquisición de este modelo de campo de luz está basada en la arquitectura óptica desarrollada por Marwah et al. (2013), donde la configuración óptica consiste en una cámara tradicional con un lente objetivo a una distancia d_a del sensor y una MCO de atenuación (MCA) instalada a una corta distancia d_m frente al sensor, como se observa en la figura 8. Específicamente, una escena $F(x,y,u,v)$, donde (x,y) corresponde a la indexación espacial y (u,v) corresponde a la indexación

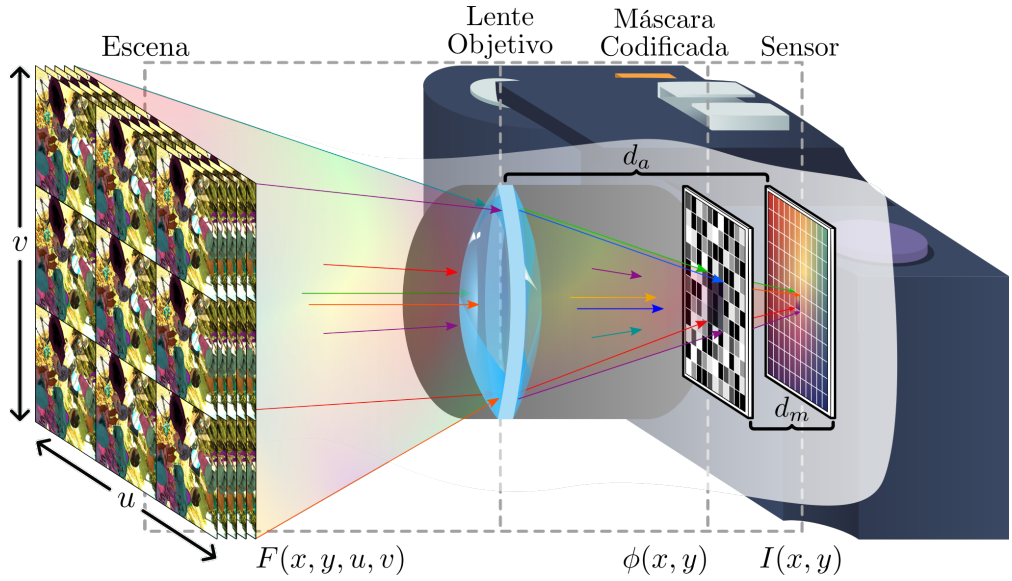
angular, es modulada por una MCA $\phi(x,y)$. Cuando el campo de luz pasa a través de la MCA, entonces es proyectado como

$$I(x,y) = \iint \phi(x+\tau(u-x), y+\tau(v-y)) F(x,y,u,v)\, du\, dv, \tag{1}$$

donde $\tau = d_m/d_a$ es el corte del patrón de la MCA respecto a la entrada de la escena. Reescribiendo la escena F como $f^s(x,y)$, donde ahora $s = 1, \cdots, |u||v|$ representa la indexación angular apilada de cada una de las vistas de F, entonces la escena será modulada por $\hat{\Phi}^s(x,y)$, donde s representa los desplazamientos angulares de ϕ respecto a la indexación angular apilada de f. En particular, asumiendo que la respuesta de la luz es invariante para cada píxel del MCA con un tamaño $\Delta_p \times \Delta_p$ Bacca et al. (2021), esta nueva representación del MCA puede ser expresada discretamente como

Figura 8

Diseño del sistema de adquisición de campos de luz con una MCA a una distancia d_m del sensor.



$$\hat{\Phi}^s(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{N} \Phi^s_{i,j}\text{rect}\left(\frac{x}{\Delta_p} - i, \frac{y}{\Delta_p} - j\right). \tag{2}$$

donde rect$(\cdot)$ es la función rectangular definida como

$$\text{rect}(x,y) = \begin{cases} 1 \quad \text{, si} \quad |x|,|y| \leq 1/2 \\ \\ 0 \quad \text{, si} \quad |x|,|y| \leq 1/2 \end{cases} \tag{3}$$

y $\Phi_{i,j}$ es el píxel en la $(i,j) - ésima$ posición espacial del $s - ésimo$ desplazamiento angular y (M,N) son la cantidad de pixeles en la MCA. Si $\Phi_{i,j} \in \{0,1\}$, esto quiere decir que la MCA es binaria. Por otro lado, $\Phi_{i,j} \in [0,1]$, entonces el modelo es más parecido a una MCA con valores reales Bacca et al. (2021). Adicionalmente, es necesario tener en cuenta que el patrón de $\Phi_{i,j}$ es único para un sistema de fotografía de campo de luz llevado a la implementación, el resto de valores proporcionados por Φ^s corresponden al desplazamiento angular apilado de forma discreta mencionado previamente.

Teniendo en cuenta las ecuaciones 1 y 2, el sistema de fotografía de campos de luz se puede modelar de forma discreta como

$$\hat{I}_{i,j} = \sum_{s=1}^{S} \Phi^s_{i,j} f^s_{i,j} + \varepsilon^s, \tag{4}$$

donde $f^s_{i,j}$ representa $(i,j) - simo$ píxel de la $S - ésima$ vista apilada de la respectiva escena, el cual corresponde al tamaño del pixel de la MCA que es modulado angularmente para obtener la $(i,j) - ésima$ muestra comprimida de la escena $\hat{I}$ y ε representa el ruido presente en el modelo.

Matemáticamente, la ecuación 4 se puede expresar sencillamente de forma vectorial como

$$\hat{\mathbf{I}} = \Phi\mathbf{f} + \hat{\varepsilon}, \tag{5}$$

donde $\hat{\mathbf{I}} \in \mathbb{R}^m$ es la medida capturada, $\mathbf{f} \in \mathbb{R}^n = \begin{bmatrix} \mathbf{f}_1^\top, & \mathbf{f}_2^\top, & \cdots, & \mathbf{f}_S^\top \end{bmatrix}^\top$ es la vectorización del campo de luz, $\Phi = \begin{bmatrix} \Phi_1, & \Phi_2, & \cdots, & \Phi_S \end{bmatrix}$ con $\Phi_i \in \mathbb{R}^{m\times m}$ siendo una submatriz de baja densidad que representa la codificación del campo de luz de la $i-ésima$ vista angular y $\hat{\varepsilon} \in \mathbb{R}^m$ representa el ruido del sistema. Finalmente, la medida comprimida de la escena completa también se puede ver como una suma ponderada de cada una de las modulaciones de las vistas angulares de la escena

$$\hat{\mathbf{I}} = \sum_{s=1}^{S} \Phi_s \mathbf{f}_s + \varepsilon_s. \tag{6}$$

3.2. Reconstrucción del Campo de Luz a partir de las Medidas Comprimidas

El tipo de reconstrucción tradicional para este tipo de medidas está basado en Marwah et al. (2013), donde el problema inverso para la reconstrucción del campo de luz requiere de la ecuación 6. Asumiendo que la escena $\hat{\mathbf{f}}$ es de baja densidad en alguna base o diccionario $\mathcal{D}^{MN\times d}$, donde d representa la cantidad de átomos presentes en el diccionario, entonces obtenemos

$$\hat{\mathbf{I}} = \Phi\mathbf{f} = \Phi\mathcal{D}\alpha, \tag{7}$$

donde la mayoría de los coeficientes $\alpha \in \mathbb{R}^d$ son cercanos a cero. La forma tradicional

del MC para realizar la reconstrucción consiste para recuperar α con la mayor cantidad de ceros posibles y que satisfaga la ecuación 7, lo que conlleva a resolver problema de optimización

$$\tilde{\mathbf{f}} = \boldsymbol{\mathcal{D}} \left(\underset{\alpha}{\text{argumento}} \ \text{mínimo} \|\hat{\mathbf{I}} - \Phi \boldsymbol{\mathcal{D}} \alpha\|_2^2 + \lambda \|\alpha\|_1 \right), \tag{8}$$

donde la norma ℓ_1 promueve la esparcidad y λ es un parámetro de regularización.

La calidad de la reconstrucción dependerá del algoritmo que sea utilizado Pope (2009), donde Φ estará directamente involucrado con el sistema físico. Cuando la recolección de reconstrucciones es usada para resolver tareas computacionales, es preferible usar las medidas comprimidas, debido a que proporcionan una mayor velocidad computacional y reducen costos de almacenamiento. Los trabajos mencionados previamente Shedligeri et al. (2017); Haim et al. (2018); Chang and Wetzstein (2019); Wu et al. (2019) demuestran que se puede obtener óptimos resultados para estimar la profundidad desde las medidas comprimidas del sistema óptico desarrollado con una matriz de muestreo Φ optimizada.

Por otra parte, mediante las metodologías del aprendizaje profundo surge una nueva forma de reconstruir los campos de luz, donde el diseño de las redes neuronales permite inferir sobre las reconstrucciones de los campos de luz a partir de las medidas comprimidas. Por mencionar algunos, Gupta et al. Gupta et al. (2017) proponen una red neuronal que se divide en dos ramas de capas completamente conectadas y capas convolucionales, respectivamente, las cuales terminan fusionándose para generar el campo de luz reconstruido. Nabati et al Nabati et al. (2018) proponen una reconstrucción a color modulando la información de color y angular mediante el uso capas

convolucionales con dilataciones. Sin embargo, no realizan el aprendizaje de la matriz de muestreo por lo cual no puede inferir de mejor manera la información de los campos de luz.

Recientemente, el método CLAP desarrollado por Guo et al. (2020) es un modelo que combina el aprendizaje de CAs junto con un decodificador tratado como un regularizador espacial y angular profundo para la reconstrucción de campos de luz como un problema inverso con un término de regularización implícito. Teniendo en cuenta una variable auxiliar $\{\mathbf{v}^s\}_{s=1}^{S} \in \mathbb{R}^n$, el problema de reconstrucción del campo de luz puede ser expresado como

$$\underset{\mathbf{x},\mathbf{v}}{\text{minimizar}} \ \frac{1}{2}\|\hat{\mathbf{I}} - \Phi\mathbf{f}\|_2^2 + \gamma\|\mathbf{f} - \mathbf{v}\|_2^2 + \lambda\mathcal{G}(\mathbf{f}), \tag{9}$$

donde $\gamma > 0$ es un parámetro penalizador, y $\mathcal{G}(\cdot)$ es un término regularizador. La ecuación 9 puede ser resuelta mediante el método de división semi cuadrática Dong et al. (2018) alternando el siguiente conjunto de problemas hasta lograr la convergencia

$$\begin{cases} \mathbf{f}^{(t+1)} = \underset{\mathbf{f}}{\text{argumento}} \ \ \text{mínimo} \ \ \frac{1}{2}\|\hat{\mathbf{I}} - \Phi\mathbf{f}\|_2^2 + \gamma\|\mathbf{f} - \mathbf{v}^{(t)}\|_2^2, \\ \mathbf{v}^{(t+1)} = \underset{\mathbf{v}}{\text{argumento}} \ \ \text{mínimo} \ \ \gamma\|\mathbf{f}^{(t+1)} - \mathbf{v}\|_2^2 + \lambda\mathcal{G}(\mathbf{f}^{(t+1)}). \end{cases} \tag{10}$$

Especificamente, la proyección de los campos de luz se puede expresar como una capa convolucional $\mathcal{P}(\cdot)$ que realiza el mapeo lineal $\mathbf{I} = \mathcal{P}(\mathbf{f})$, y la proyección inversa se puede expresar como una capa deconvolucional $\mathcal{R}(\cdot)$ que realiza el mapeo $\mathbf{f} = \mathcal{P}(\mathbf{I})$. Por lo tanto, el problema de optimización se puede expresar como

$$\begin{cases} \mathbf{v}^{(t+1)} = \boldsymbol{\mathcal{D}}(\mathbf{f}^{(t)}, \theta_d^t), \\ \mathbf{f}^{(t+1)} = \mathbf{f}^{(t)} - \delta_t[\boldsymbol{\mathcal{R}}(P(\mathbf{f}^{(t)}, \theta_p^t) - \mathbf{I}, \theta_r^t) + \gamma_t(\mathbf{f}^{(t)} - \mathbf{v}^{(t+1)})], \end{cases} \tag{11}$$

donde $\boldsymbol{\mathcal{D}}(\cdot)$ es un operador proximal respecto al regulizador $\boldsymbol{\mathcal{G}}(\cdot)$, y θ_r, θ_p y θ_d son los parámetros de la red.

4. Estimación de la Profundidad del Campo de Luz Comprimido

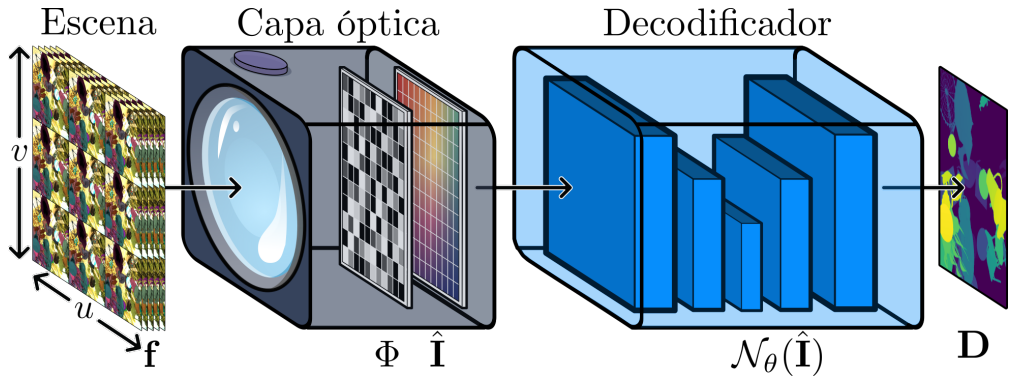
Como se observó en el capítulo 2, los métodos actuales para estimar la profundidad se basan exclusivamente en campos de luz completos o reconstruidos. En este capítulo se explicará el método propuesto para realizar una estimación de la profundidad mediante un enfoque EAE, como se observa en la figura 9. Cabe resaltar que dada la implícita relación entre los mapas de disparidad y profundidad, el estado del arte contra el cual se comparará se enfoca en la estimación de mapas de disparidad, por lo tanto, el método propuesto se enfocará principalmente en estimar mapas de disparidad.

4.1. Estimación de la profundidad de Extremo a Extremo

El método propuesto se divide en dos etapas: la etapa de entrenamiento donde se optimiza la máscara codificada y los parámetros de la RNC utilizando las muestras de entrenamiento y una

Figura 9

Método propuesto con enfoque de extremo a extremo.



Nota: Se divide en dos capas: La capa óptica, genera las medidas comprimidas a partir de la modulación espacial y angular de las escenas mediante la MCA. El decodificador se enfoca en estimar los mapas de disparidad a partir de las medidas comprimidas.

etapa de inferencia que consiste en evaluar el modelo propuesto con muestras que no han sido empleadas en el entrenamiento.

4.1.1. Étapa de entrenamiento. En esta etapa, se diseña una red neuronal que consta de dos capas principales, la capa óptica cuyo objetivo es aprender un muestreo óptimo para el campo de luz basado en MC, seguido del decodificador que busca aprender los parámetros óptimos de una RNC para estimar los mapas de disparidad. Matemáticamente, si de un conjunto de escenas $\{\mathbf{f}_i\}_{i=1}^{M}$ con sus respectivos mapas de disparidad verdaderos $\{\mathbf{D}_i\}_{i=1}^{M}$, las escenas son moduladas por la capa óptica Φ, generando medidas comprimidas $\hat{\mathbf{I}}_i = \Phi\mathbf{f}_i$, donde los parámetros de Φ son entrenables, entonces estas medidas generarán los mapas de disparidad al ser procesadas por el decodificador $\mathcal{N}_\theta$, de tal manera que la estimación será $\hat{\mathbf{D}}_i = \mathcal{N}_\theta(\Phi\mathbf{f}_i)$, donde los parámetros θ también son entrenables. Para resolver este problema, se desarrolla una optimización conjunta de los parámetros Φ del modelo de muestreo y los parámetros θ del modelo de RNC, de tal manera que el problema puede ser expresado como

$$\begin{aligned} \{\Phi^*, \theta^*\} = \quad & \underset{\Phi,\theta}{\text{argumento mínimo}} \quad \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}(\mathcal{N}_\theta(\Phi\mathbf{f}_i), \mathbf{D}_i) \\ & \text{subject to} \qquad\qquad \Phi_{ii} \in [0,1], \quad 1 \leq i \leq m, \quad 0 \leq j \leq m. \end{aligned} \tag{12}$$

$\{\mathbf{D}_i\}_{i=1}^{M}$ corresponde al conjunto de mapas de disparidad verdaderos del respectivo conjunto de escenas

donde Φ^* es la MCA optimizada para resolver esta tarea y θ^* son los parámetros óptimos

de $\mathcal{N}_\theta$. Se debe tener en cuenta que las restricciones de la ecuación 12 están basadas en un sistema que puede ser implementable en la vida real. Esta capa óptica es conectada al decodificador como se muestra en la figura 9. Se puede apreciar que el diseño de la MCA afecta directamente al modelo de RNC debido a que todos sus pesos son actualizados junto con los mismos valores de la MCA durante la propagación hacia atrás del gradiente. Debido a la naturaleza del método propuesto, las dos capas se pueden separar para que desarrollen las tareas específicas, donde la capa óptica permitiría adquirir escenas con MCAs optimizadas para estimar mapas de disparidad, y el decodificador puede funcionar como una red preentrenada para interpretar nuevos datos codificados y obtener los mapas de disparidad. En la sección 4.2 se explica en detalle una RNC basada en la arquitectura U-net Ronneberger et al. (2015) con bloques residuales He et al. (2016) y proyecciones ascendentes Laina et al. (2016). Cabe resaltar que el decodificador puede ser cualquier tipo de RNC que transforme las muestras de entrada en las muestras de salida deseadas.

4.1.2. Étapa de Inferencia. Una vez terminada la etapa de entrenamiento, la capa óptica y el decodificador podrán realizar sus tareas específicas de forma óptima. Específicamente, la capa óptica obtendrá las medidas comprimidas optimizadas del conjunto de escenas entrenadas de tal manera que

$$\hat{\mathbf{I}}^* = \Phi^* \mathbf{f}_r, \tag{13}$$

donde $\mathbf{f}_r$ son muestras del mundo real o para evaluar el algoritmo y Φ^* representa la MCA optimizada que puede ser implementada en hardware, mientras que el decodificador podrá estimar

los mapas de disparidad $\hat{\mathbf{D}}$ a partir de estas medidas optimizadas, que puede ser expresado como

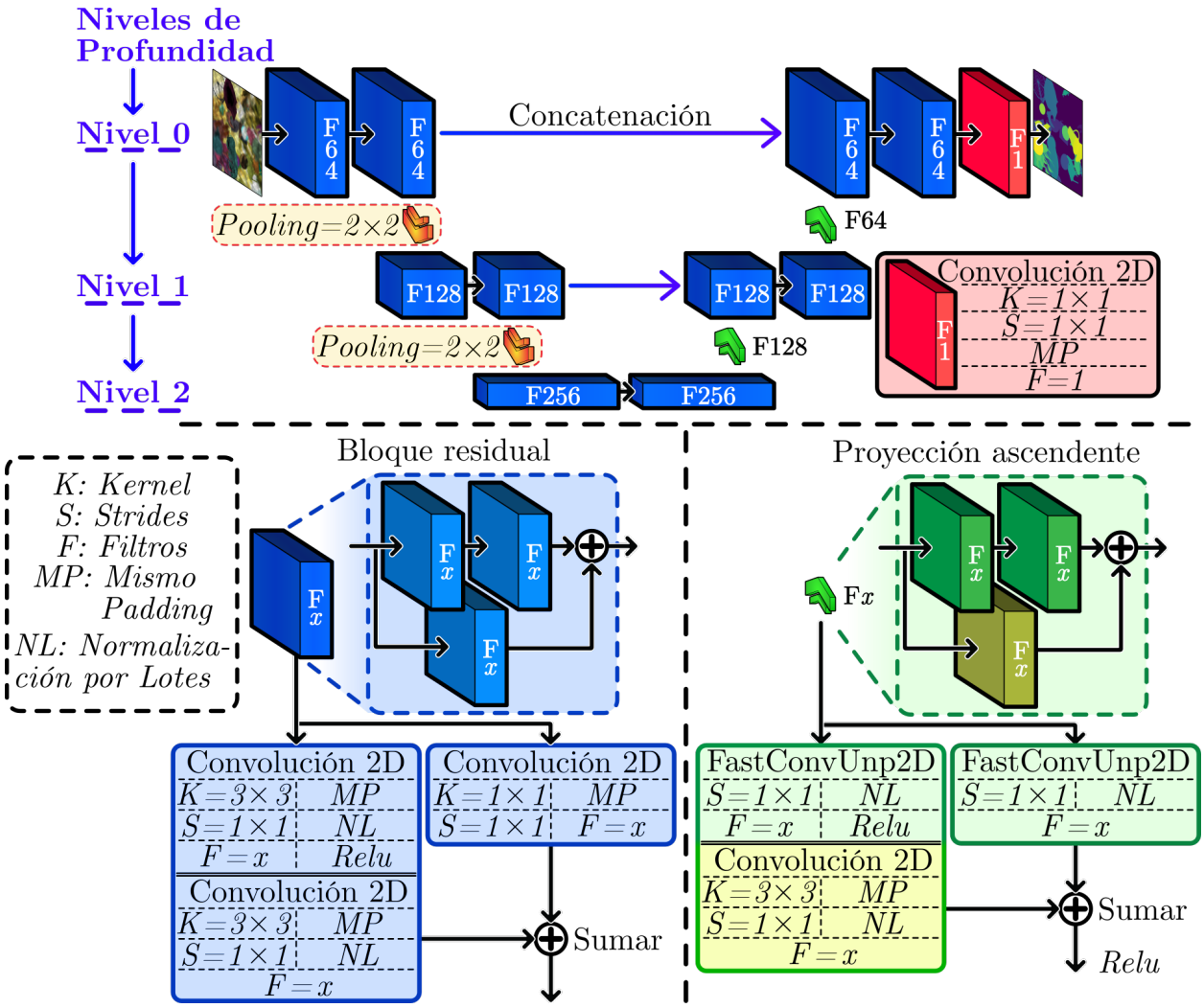
$$\hat{\mathbf{D}} = \mathcal{N}_{\theta^*}(\hat{\mathbf{I}}^*) = \mathcal{N}_{\theta^*}(\Phi^*\mathbf{f}). \tag{14}$$

4.2. Decodificador Profundo

La principal motivación del modelo propuesto como decodificador en este trabajo surge debido a que las muestras comprimidas para el entrenamiento son pocas en comparación con muestras de campos de luz completos y, se encuentran en un dominio escaso dificultando el entrenamiento para las RNCs tradicionales Li et al. (2021b), especialmente cuando está centrada en resolver alguna tarea computacional distinta a la reconstrucción Bacca et al. (2021). Este modelo se encuentra fuertemente inspirado en las arquitecturas propuestas por Laina et al. (2016) y Harsányi et al. (2018). La estructura del decodificador del método propuesto se divide en 3 partes principales: un codificador que se encarga de aprender el mapeo de características más representativas de la entrada de datos; un cuello de botella, donde el modelo aprende la representación de los datos; y un decodificador que realiza convoluciones y muestreos ascendentes de la información codificada hasta la dimensionalidad deseada. La mayor dificultad para realizar esta tarea consiste en extraer características de las medidas codificadas debido a que difieren de las típicas características que una RNC podría extraer fácilmente desde el codificador, como lo son líneas, curvas y bordes. Esto genera el problema de reutilización de características decrecientes, donde el mapeo de características se pierde en el cálculo de las primeras capas convolucionales Zagoruyko and Komodakis (2016). Para resolver este problema se propone el modelo de la figura 10. A continuación, se dará

Figura 10

Modelo para el decodificador del método propuesto.



Nota: Modelo basado en la arquitectura U-net amplia donde cada capa de la red consiste en bloques residuales, muestreo descendente MaxPooling y proyecciones ascendentes.

la descripción de las partes del modelo y los aportes que realiza a la tarea en general.

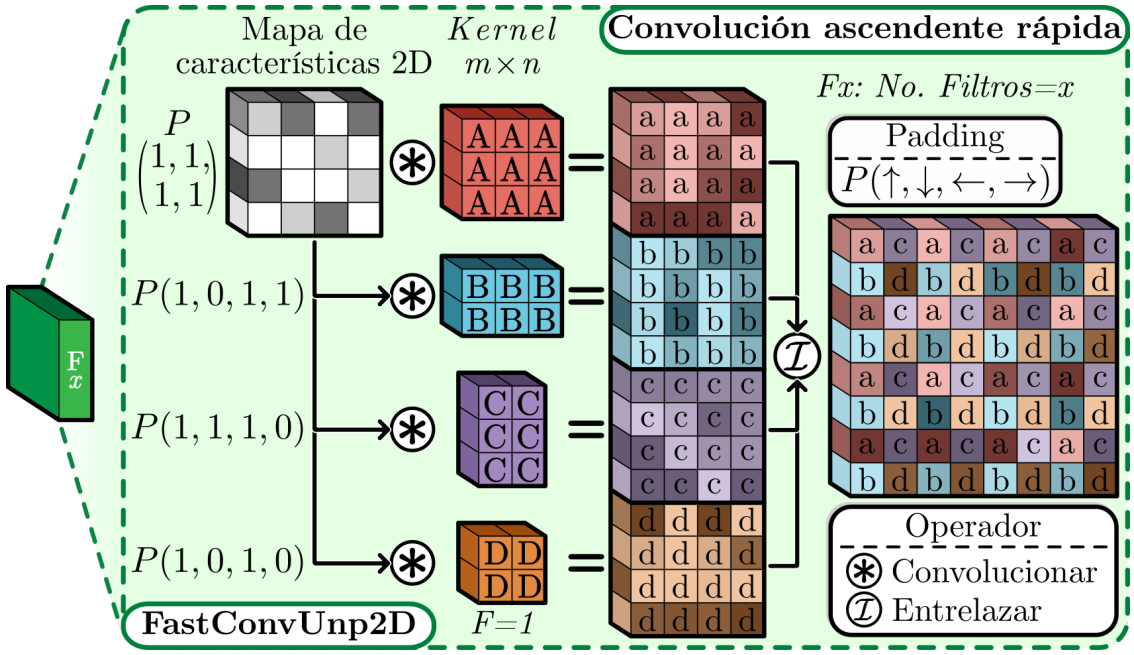
- **Estructura:** variación de la red U-net Ronneberger et al. (2015) con un enfoque de red amplia Zagoruyko and Komodakis (2016) cuyo objetivo es mitigar el problema de reutilización de características decrecientes, donde la entrada generada por la capa óptica se desvanece

durante el mapeo de características. Al disminuir la profundidad de las redes neuronales y en cambio, agregar más filtros a las capas convolucionales se puede resolver experimentalmente este problema conocido como reutilización de características decrecientes. Por lo tanto, se mantiene la misma intuición para el modelo propuesto, que solo contiene 2 niveles de profundidad a comparación de la red Harsányi et al. (2018) que contiene 4 niveles de profundidad, y con una cantidad elevada de filtros convolucionales por bloque residual.

- **Bloque residual:** es el bloque base de este modelo. Como se observa en la figura 10, este bloque se encuentra representado en color azul oscuro, y está constituido por 3 capas convo-

Figura 11

Convolución ascendente rápida



Nota: Propuesta por Laina et al. (2016). Esta es una técnica de muestreo ascendente donde se evita expandir espacialmente el mapa de características con valores cero como en los métodos tradicionales mencionados en 2.3.2

lucionales distintas. 2 capas operan la entrada de forma simultánea, y la otra capa que opera nuevamente la entrada, para adaptarla a la misma cantidad de filtros de las otras dos capas convolucionales para finalmente general la salida deseada. Como se explicó en la subsección 2.3.3, los bloques residuales mitigan el problema del desvanecimiento del gradiente, permitiendo un mejor entrenamiento de los pesos en todo el modelo.

- **Proyección ascendente:** es el muestreo ascendente del modelo, basado en Laina et al. (2016). Este tipo de muestreo tiene una estructura externa idéntica al bloque residual 10. La diferencia subyace en la convolución ascendente rápida representada por el bloque de color verde. Como se observa en la figura 11, dado un mapa de características, se le aplican 4 operaciones convolucionales con 4 diferentes kernels, donde $m \times n$ representa la cantidad de filas y columnas presentes. Estas convoluciones son entrelazadas para formar el nuevo mapa de características donde cada letra en minúscula representa el bloque convolucional de donde proviene. Cabe resaltar si se usarán las 4 convoluciones de forma tradicional, representaría una única capa convolucional con kernel 5×5, y la cantidad de filtros puede ser variable. Esta operación permite ampliar espacialmente los mapas de características sin perder información en el proceso.

- **Muestreo descendente:** consiste en dos bloques de MaxPooling 2×2, como se mencionó en 2.3.2. Este muestreo descendente permite retener los pixeles de información más relevantes hasta llegar al cuello de botella. Diferente de AvgPooling 2×2, que promedia todas estas regiones, ocasionando una perdida significativa de la información escasa inicial.

- **Convolución final:** está representado por el bloque rojo en la figura 10, cuya función es generar el mapa de características final que representaría al mapa de disparidad estimado.

En resumen, esta RNC tiene un óptimo rendimiento para entradas de datos dispersas basadas en MC porque mitiga dos problemas de las RNCs: el desvanecimiento del gradiente y la disminución de la reutilización de características. Además, aprovecha un muestreo ascendente que mejora la conservación de los mapas de caracteristicas durante la decodificación.

5. Simulaciones y Resultados

En este capítulo se presentará el conjunto de datos utilizado, junto con la respectiva adquisición de las medidas comprimidas. Se presentarán las principales métricas de evaluación tanto para el método propuesto como para realizar las respectivas comparaciones. Y finalmente, se presentarán todas las configuraciones realizadas con los métodos de comparación, junto con los respectivos análisis de resultados.

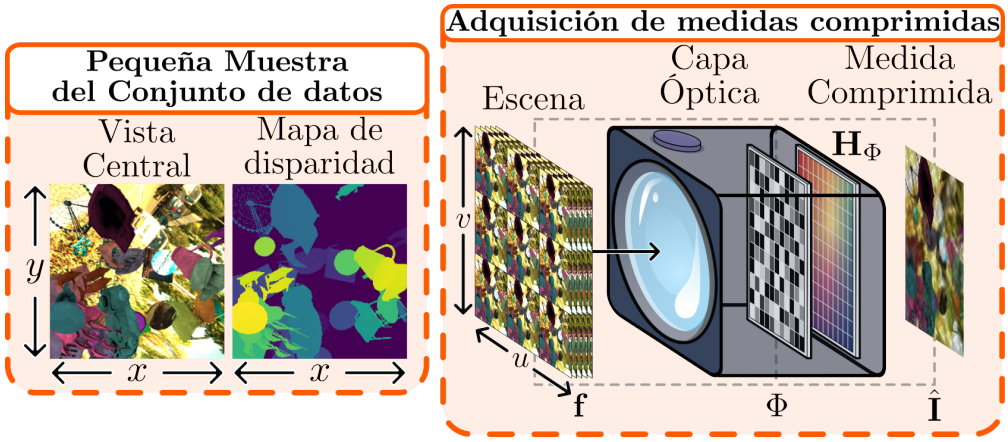
5.1. Conjunto de Datos

El conjunto de datos usado en este trabajo fue elaborado por Schambach and Heizmann (2020), consta de 500 escenas sintéticas de campos de luz multiespectrales con una precisión numérica de entero sin signo de 16 bits con sus respectivos campos de disparidad, como se observa en la parte izquierda de la figura 12. Cada escena está conformada por un campo de luz con resolución angular de 11×11, una resolución espacial de 512×512 y con 13 bandas espectrales. De los campos de luz espectrales se trabajó solo con la información RGB que fue obtenida usando las funciones de combinación de colores CIE 1931 y el iluminante de luz de mediodía CIE D65. El conjunto de datos está dividido en 3 partes: 400 muestras para entrenamiento, 50 muestras para validación y 50 muestras para prueba.

Para reducir el costo computacional, todas las muestras fueron reducidas angular y espacialmente a 7×7 y 256×256, respectivamente. Para los campos de disparidad se tomaron únicamente sus vistas centrales. Para un correcto entrenamiento del modelo propuesto de los datos, los campos de luz como mapas de disparidad fueron normalizados mediante $z_i = \frac{x_i - min(x)}{max(x) - min(x)}$, donde x_i es la

Figura 12

Muestras usadas en las configuraciones.



Nota: Parte izquierda: Muestra de la vista central de un campo de luz sintético con resolución espacial de 256×256 *con su respectivo mapa de disparidad. Parte derecha: Muestra del campo de luz comprimido con resolución espacial* 256×256 *obtenida de la modulación del campo de luz con resolución angular* 7×7 *y resolución espacial* 256×256.

muestra a normalizar del conjunto de datos x, ya sean los campos de luz o mapas de disparidad. Posteriormente, cada muestra fue recortada en parches con la misma resolución angular y una resolución espacial de 32×32, por lo que se generaron 64 parches por muestra para un total de 32000 parches por todo el conjunto de datos.

Como se observa en la parte derecha de la figura 12, para obtener las medidas comprimidas se hizo usó de la teoría mencionada en la sección 3, donde se usaron dos tipos de máscara de atenuación: una máscara completamente aleatoria y otro máscara optimizada mediante el método propuesto. El sensado solo fue aplicado a los campos de luz en formato RGB y las nuevas muestras adquiridas solo tienen la dimensión espacial de 32×32.

Finalmente, se realizó un aumento de los datos para el entrenamiento de las redes neuronales que consistió en girar horizontalmente las muestras, rotar las muestras $0°, 90°, 180°$ y $270°$, y se

aplicó un estiramiento gamma a los canales de los campos de luz. El aumento de datos se aplica de forma aleatoria cada mini lote de muestras durante el entrenamiento, generando una mejor aleatoriedad de las muestras. En total se generaron 25600 parches para entrenamiento y 3200 parches para validación y prueba, respectivamente.

5.2. Función de Pérdida y Métricas

Para evaluar cuantitativamente los resultados obtenidos, se usaron diversas métricas que permiten cuantificar de forma precisa las diferencias de píxel a píxel entre los mapas de disparidad verdaderos y sus predicciones. Para entrenar y evaluar el método propuesto se usó la función de pérdida de pseudo Huber

$$\mathcal{L}_\delta(e_i) = \delta^2 \left(\sqrt{1 + \left(\frac{e_i}{\delta}\right)^2} - 1 \right) \tag{15}$$

donde $e_i = |\mathbf{d}_i - \hat{\mathbf{d}}_i|$ es el error absoluto del $i-ésimo$ de un mapa de disparidad verdadero vectorizado $\mathbf{d} \in \mathbb{R}^m$ con respecto a su reconstrucción $\hat{\mathbf{d}} \in \mathbb{R}^m$, y la pérdida total es estimada como la media de cada uno de los e_i estimados. Esta pérdida es una aproximación suave de la función de pérdida de Huber Huber (1992), donde δ permite controlar los valores atípicos. para pequeños valores de e_i la función de pseudo Huber se aproxima al error cuadrático medio y para grandes valores de e_i se aproxima a una recta con pendiente δ. Finalmente, esta función se convierte fuertemente convexa cuando se alcanzan valores óptimos.

Las otras métricas usadas para evaluar la calidad de las predicciones son el error cuadrático medio (MSE)

$$MSE = \frac{1}{M}\sum_{i=1}^{M} e_i^2, \tag{16}$$

y el error absoluto medio (MAE)

$$MAE = \frac{1}{M}\sum_{i=1}^{M} e_i, \tag{17}$$

los cuales permiten verificar la consistencia de la predicción de forma independiente, es decir, para valores esperados como valores atípicos.

Y finalmente, la mala relación de pixeles (BadPix), expresada como

$$BadPix(t) = 100 * \sum_{i=1}^{m} \frac{|e_i| > t}{M} \tag{18}$$

donde se calcula la cantidad de pixeles que se desvían más allá del valor t de los mapas de disparidad verdaderos, y M representa la cantidad de pixeles en cada muestra.

5.3. Experimentos

En esta subsección se estudian los resultados obtenidos para estimar la disparidad mediante el método propuesto y se da una descripción de los métodos implementados para comparación. Se muestran y analizan los resultados obtenidos sobre la estimación de la disparidad. Finalmente, se analizan las MCAs usadas y obtenidas desde el método propuesto.

5.3.1. Configuraciones del Método propuesto. El método propuesto será evaluado bajo dos configuraciones distintas: una configuración donde se entrena el modelo propuesto

EAE y otra donde la capa óptica del modelo propuesto tiene una MCA aleatoria, por lo que solo se entrenará el decodificador. La motivación de estas configuraciones consiste en comparar la velocidad de convergencia dado que la configuración RNC entrena la red neuronal unicamente con las muestras comprimidas, por lo que el costo de almacenamiento se reduce a $\frac{1}{7 \cdot 7} = \frac{1}{49}$ veces su tamaño original y aumenta la velocidad de entrenamiento, permitiendo procesar estos datos en equipos con poco espacio del almacenamiento. Por otra parte, también se puede medir la pérdida de calidad entre estas dos configuraciones dado que la configuración RNC no aprende a inferir una MCA óptima para el conjunto de datos como si lo hace la configuración EAE.

Para el entrenamiento del método propuesto para ambas configuraciones se realizaron 100 épocas, con mini lotes de 16 muestras, para un total de 1600 mini lotes por época. Se usó el optimizador de propagación cuadrática media (RMSprop) con una tasa de aprendizaje fijo de $5e-4$.

5.3.2. Métodos de Comparación. Los métodos de comparación empleados consisten en estimar la profundidad mediante los modelos Epinet Shin et al. (2018) y Lfattnet Tsai et al. (2020) usando campos de luz reconstruidos bajo dos metodologías: algoritmos computacionales (CLMC) Marwah et al. (2013) o con redes neuronales (CLAP) Guo et al. (2020). Se establecen estos métodos de comparación debido a que no se han reportado trabajos en el estado del arte que se puedan comparar al método propuesto en este trabajo. Para realizar una adecuada comparación contra el método propuesto, estos modelos serán entrenados con los respectivos conjuntos de campos de luz reconstruidos desde cero. Es importante resaltar que las configuraciones y métodos implementados fueron entrenados con Tensorflow Abadi et al. (2016) usando una tarjeta de video

NVIDIA GeForce RTX 3090, excepto la reconstrucción basada en CLMC donde se hizo uso de MATLAB Matlab (2012) sin GPU.

A continuación, se describen cada una de las implementaciones realizadas:

Epinet: red neuronal basada en Shin et al. (2018). El aumento de datos consistió en reflejar horizontalmente las muestras, rotarlas $0°, 90°, 180°$ y $270°$, aplicar un estiramiento gamma a los canales de los campos de luz y escalar espacialmente las muestras en un pequeño rango. Para el entrenamiento se convirtieron los campos de luz a escala de grises, con mini lotes de 16 muestras, se usó el optimizador RMSprop con una tasa de aprendizaje decreciente entre $1e-5$ hasta $1e-6$ y se hicieron 200 épocas debido a que en ese punto el entramiento del modelo alcanza una tasa de convergencia estable.

Lfattnet: red neuronal basada en Tsai et al. (2020). El aumento de datos consistió en reflejar horizontalmente las muestras y rotarlas $0°, 90°, 180°$ y $270°$. Para el entrenamiento se usaron mini lotes de 12 muestras, se usó Adam, con una tasa de aprendizaje fija de $1e-3$ durante 70 épocas donde alcanza una tasa de convergencia estable. Este modelo fue adaptado para entrenar campos de luz con resolución angular 7×7, debido a que el mapa de atención que se calcula cambia de acuerdo con esta resolución. Para entrenar este modelo con el conjunto de datos generado usando Guo et al. (2020) fue necesario realizar una modificación al módulo SPP del modelo debido a que la entrada de datos para este módulo corresponde a una sola de las vistas angulares con una resolución espacial de 36×36 cuyas dimensiones no coinciden con el submuestreo e interpolación bilineal que se realiza en este módulo . Por lo tanto, la modificación que se realizó consistió en una agrupación promedio con tamaños 2×2, 4×4, 9×9 y 18×18. Después de la agrupación,

se utiliza una capa de convolución de 1×1 para reducir la dimensión del mapa de características actual. Luego se realiza un proceso de interpolación bilineal para aumentar la muestra de estos mapas de características de baja dimensión al mismo tamaño. Finalmente, se concatenaron los mapas de características de todos los niveles.

CLMC: reconstrucción de campos de luz basada en Marwah et al. (2013). Para esta reconstrucción se debe aprender un diccionario. En este caso, se aprendió un diccionario $1\times$ sobrecompleto, para una MCA con resolución angular 7×7 y resolución espacial 16×16. Las muestras fueron tomadas aleatoriamente del conjunto de datos para entrenamiento y se convirtieron a escala de grises. Se realizó el proceso de aprendizaje con el algoritmo KSVD Aharon et al. (2006) durante 10 iteraciones. En la reconstrucción de los campos de luz se usó el algoritmo ADMM Boyd et al. (2011), con el parámetro regularizador $\lambda = 1e-4$, el parámetro lagrangiano aumentado $\rho = 1e-2$ y el parámetro de relajación excesiva $\alpha = 1$. Las reconstrucciones fueron aplicadas para cada banda de los campos de luz para obtener la reconstrucción final de los campos de luz a color.

CLAP: reconstrucción de campos de luz basada en Guo et al. (2020). Para realizar esta re-

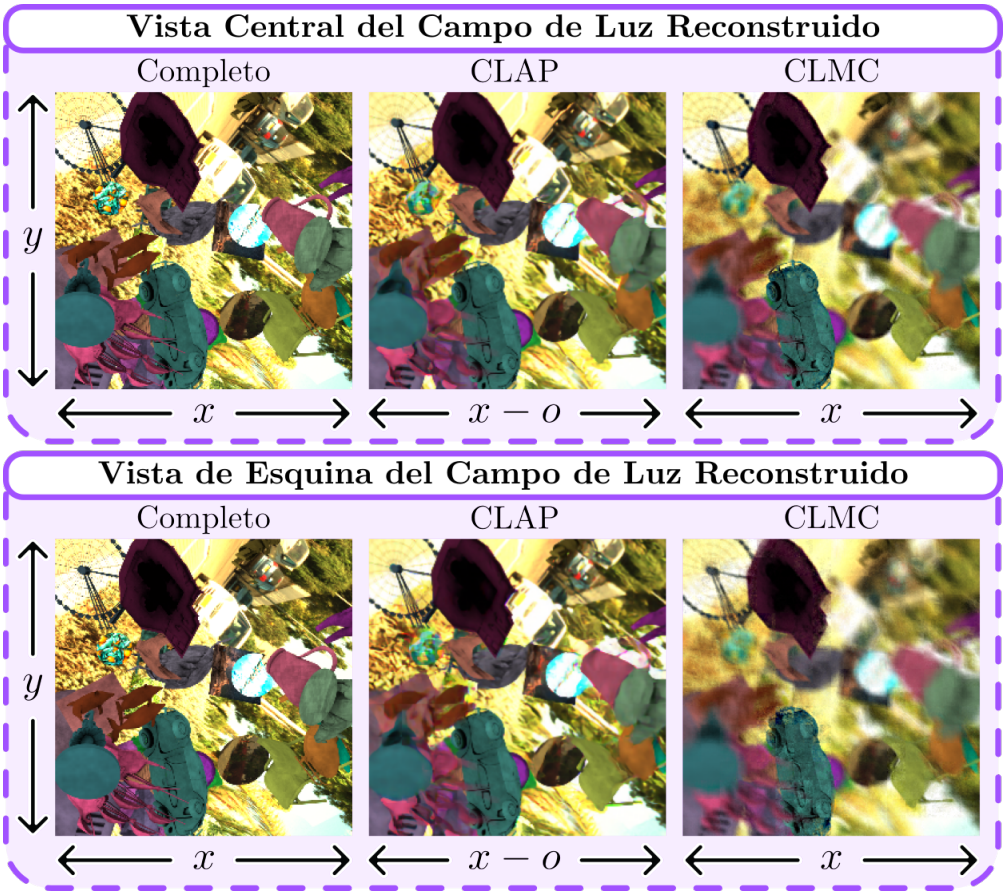
Tabla 1

Resultados cuantitativos de las reconstrucciones de los campos de luz mediante MC y aprendizaje profundo junto con sus tiempos de entrenamiento e inferencia.

Modelo	Pseudo huber	MAE	MSE	PSNR	Tiempo	
					Entrenamiento	Inferencia
CLMC	0.0063	0.0641	0.0119	20.5282	8.166 horas	30 minutos
CLAP	0.0029	0.0357	0.0053	24.3576	3.82 días	6.5 minutos

Figura 13

Reconstrucciones de campos de luz.



Nota: Muestra de vista central y muestra de vista ubicada en una esquina del campo de luz con las respectivas reconstrucciones CLMC y CLAP. Completo: muestra sintética. CLAP: reconstrucción mediante aprendizaje profundo propuesto por Guo et al. (2020), CLMC: reconstrucción mediante MC propuesto por Marwah et al. (2013).

construcción, los autores entrenan el CA como si fuera una capa convolucional usada para realizar la operación de compresión de los campos de luz como para realizar la operación transpuesta, que consiste en la reconstrucción. En este caso, se modificó el sistema de adquisición por el empleado en el sistema óptico propuesto. Este modelo fue entrenado con muestras de campos de luz en escala de grises con resolución angular de 7×7 y resolución espacial de 32×32, se usa la función de

pérdida MAE, y el optimizador Adam con una tasa de aprendizaje decreciente que inicia en $1e-4$ y se reduce a $1e-5$. El entrenamiento fue realizado durante 100 épocas y tomó casi 4 días en realizarse como se observa en la tabla 1. Similar a CLMC, las reconstrucciones de los campos de luz se hacen por cada una de las bandas espectrales para formar los campos de luz a color. Para esta reconstrucción se recortaron espacialmente los bordes de cada parche de campo de luz reconstruido para reducir la cantidad de artefactos presentes en la reconstrucción final, generando campos de luz reconstruidos completos con resolución angular de 7×7 y resolución espacial de 252×252. Para entrenar los modelos de comparación se extrajeron parches con la misma resolución angular y una nueva resolución espacial de 36×36.

Al comparar las reconstrucciones realizadas se puede ver que la calidad de reconstrucción de CLAP es superior a CLMC, como se observa en la figura 13 y en la tabla 1. Adicionalmente, se observa el tiempo total de ambos métodos donde CLAP toma más de 11 veces el tiempo en aprender a realizar las reconstrucciones que CLMC. Sin embargo, su tiempo de inferencia es 4.6 veces más rápido que CLMC. Para CLAP, $x-o$ representa el ancho de la resolución angular menos el recorte de los bordes realizado para reducir los artefactos generados por la reconstrucción por parches. En la parte inferior de la misma figura se muestra la reconstrucción de una de las vistas que se encuentran ubicadas en la esquina o borde del campo de luz, se observa que CLMC obtiene una menor calidad visual respecto a su vista central reconstruida en comparación con CLAP. Dado que los campos de luz verdaderos se encuentran normalizados en el rango $[0,1]$ para el entrenamiento de los métodos CLMC y CLAP, entonces sus respectivas reconstrucciones también están normalizadas, excepto por valores atípicos que se desbordan, lo cual puede generar problemas al

momento de usarlos para entrenar una RNC. Por lo tanto, los valores que se desbordan son proyectados, de tal manera que los valores $x > 1$ son proyectados a $x = 1$. Igualmente, si $x < 0$, entonces son proyectados a $x = 0$.

5.3.3. Resultados. En la Figura 14 se observa un subconjunto de muestra de los campos de luz completos y comprimidos, así como los mapas de disparidad que serán empleados en el entrenamiento. Cabe resaltar que si se observan las muestras de izquierda a derecha, se visualiza la pérdida de información angular y espacial.

La evaluación cuantitativa de los mapas de disparidad estimados con el método propuesto y los métodos de comparación se reportan en la tabla 2, donde se realiza la comparación entre los modelos propuestos entrenados con los respectivos conjuntos de datos desde cero. Los resultados anexados a esta tabla corresponden al mejor rendimiento de cada una de las configuraciones y mé-

Tabla 2

Resultados cuantitativos de la estimación de los mapas de disparidad.

Configuraciones y Métodos	Pseudo Hubber	MAE	MSE	BadPix01	BadPix03	BadPix07	Tiempo	
							Entrenamiento (horas)	Inferencia (minutos)
Referencia	0.0003	0.0049	0.0005	4.6220	3.0981	2.0559	110	7
Epinet + CLMC	0.0050	0.0525	0.0096	63.4084	44.5679	23.3795	*20.944*	30.16
Epinet + CLAP	*0.0018*	*0.0231*	*0.0031*	*42.3392*	*16.7245*	*7.6568*	104.457	*6.78*
Lfattnet + CLMC	0.0033	0.0363	0.0058	50.3248	27.7073	14.6838	26.638	30.45
Lfattnet + CLAP	0.0014	**0.0178**	0.0024	**30.2282**	**11.1600**	6.1042	110.512	7.61
Config. CNN	0.0034	0.0458	0.0057	73.8825	43.4797	19.1005	**2.639**	**0.05**
Config. E2E	**0.0013**	0.0189	**0.0020**	38.3156	13.9681	**5.6850**	5.416	0.33

Nota: En esta tabla se muestra el resultado promedio obtenido para cada una de las configuraciones como métodos de comparación junto con los tiempos de entrenamiento e inferencia.

todos de comparación implementados. La referencia en la tabla se refiere al rendimiento más alto de los modelos usando las muestras de campos de luz completos que correspondió al entrenamiento y evaluación del modelo Lfattnet, por lo tanto, no será tenido en cuenta para comparar contra el método propuesto. En la tabla se observa que el rendimiento del método propuesto mediante la configuración EAE es superior en las métricas de $Pseudo - Hubber$, MSE y $BadPix07$ al modelo Lfattnet entrenado con CLAP, pero inferior en las demás métricas. Por lo tanto, la calidad de las estimaciones de los mapas de disparidad son comparables al método de comparación Lfattnet entrenado con CLAP que se encuentra por debajo de ambos.

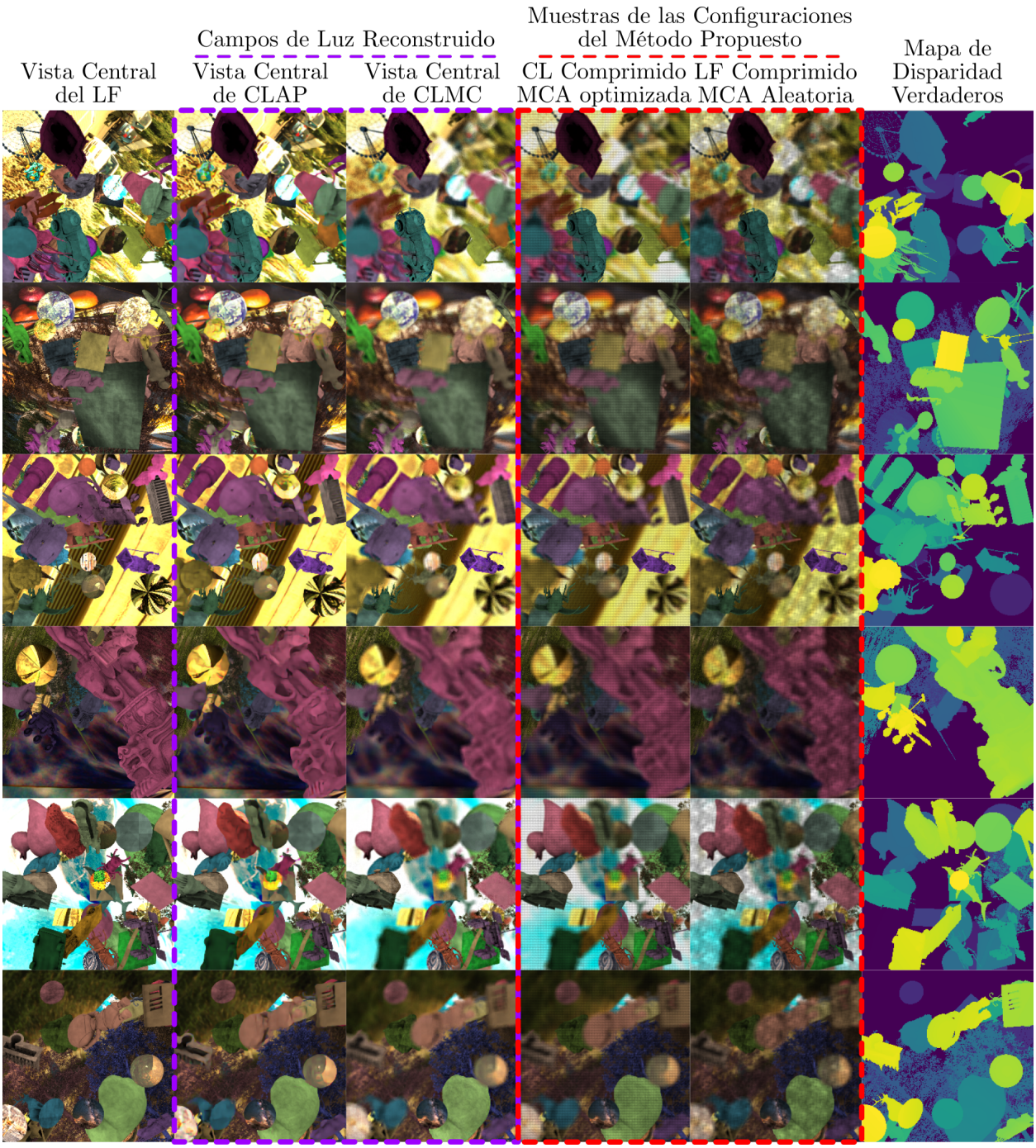
La mayor ganancia del método propuesto se encuentra en el tiempo de convergencia, tanto para entrenamiento como para inferencia, donde la configuración EAE tarda 195 segundos en entrenar una época. Por lo tanto, para 100 épocas tarda un tiempo total de 5,416 horas de entrenamiento mientras que Lfattnet entrenado con CLAP tarda un total de 18,5 horas siendo 3,4 veces más rápido la configuración EAE. Además, teniendo en cuenta que el tiempo de la reconstrucción tomó aproximadamente 4 días, el método propuesto es 20 veces más rápido que el mejor resultado obtenido por los modelos de comparación. Aplicando el mismo análisis a Epinet, la configuración EAE es 2,3 veces más rápido y teniendo en cuenta CLAP, sería 19 veces más rápido. Se puede apreciar que el entrenamiento con la configuración RNC es la más rápida en ser realizada, seguida de la configuración EAE, Epinet y finalmente, Lfattnet. Al analizar los tiempos de inferencia se observa que ambas configuraciones del método propuesto son más rápidas en este proceso que todos los métodos de comparación. Especificamente, la configuración EAE logra ser 23 veces más rápido que el mejor método de comparación.

Los resultados visuales se observan en la figura 15, donde se puede apreciar todas las estimaciones realizadas. Al observar las estimaciones realizadas mediante las configuraciones del método propuesto se puede concluir que la configuración EAE es capaz de estimar detalles finos como bordes donde visualmente se presenta borroso debido a la compresión como en los objetos del fondo a comparación de los modelos que usan los campos de luz reconstruidos. Por otro lado, las estimaciones a partir de CLMC contienen demasiados artefactos, sin embargo se mantienen comparables únicamente a la configuración RNC como se observa en la tabla 2, pero se aprecia la diferencia de que la configuración RNC reduce la cantidad de artefactos presentes, dando una visualización más coherente de los mapas de disparidad.

La ACM aleatoria usada para la configuración RNC y la ACM optimizada mediante el enfoque EAE se observa en la figura 16. El tamaño de las AMCs es de tamaño real en píxel con una resolución de 32×32. Se observa que la ACM optimizada aprende el patrón de cuadricula que es óptimo para estimar la profundidad mediante el decodificador profundo empleado.

Figura 14

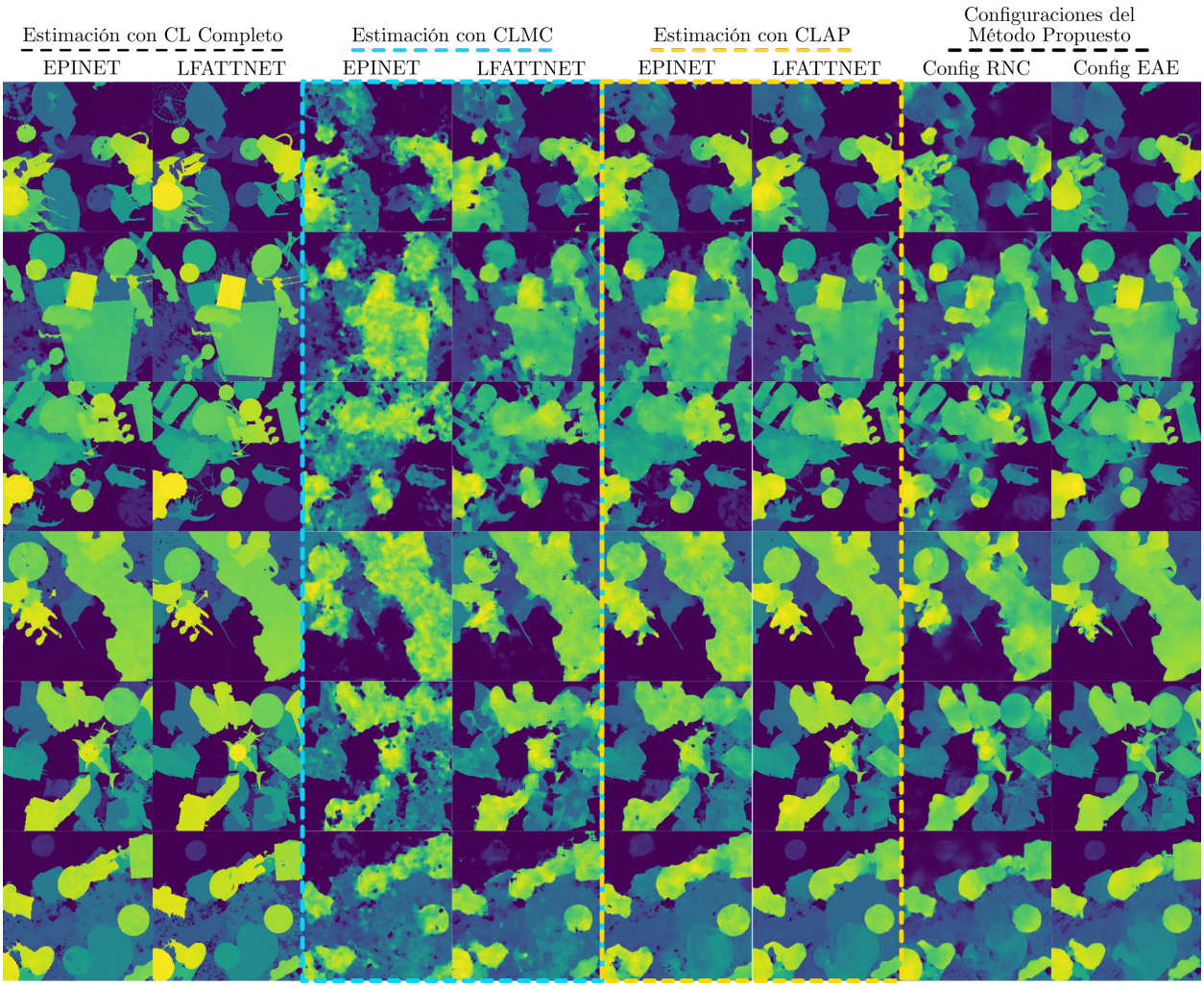
Muestras de campos de luz completas y comprimidas.



Nota: Bloque morado: Vista central de las reconstrucciones de campos de luz CLMC y CLAP. Bloque rojo: Muestras comprimidas para las configuraciones RNC y EAE. Las muestras a color están normalizadas por motivos de visualización.

Figura 15

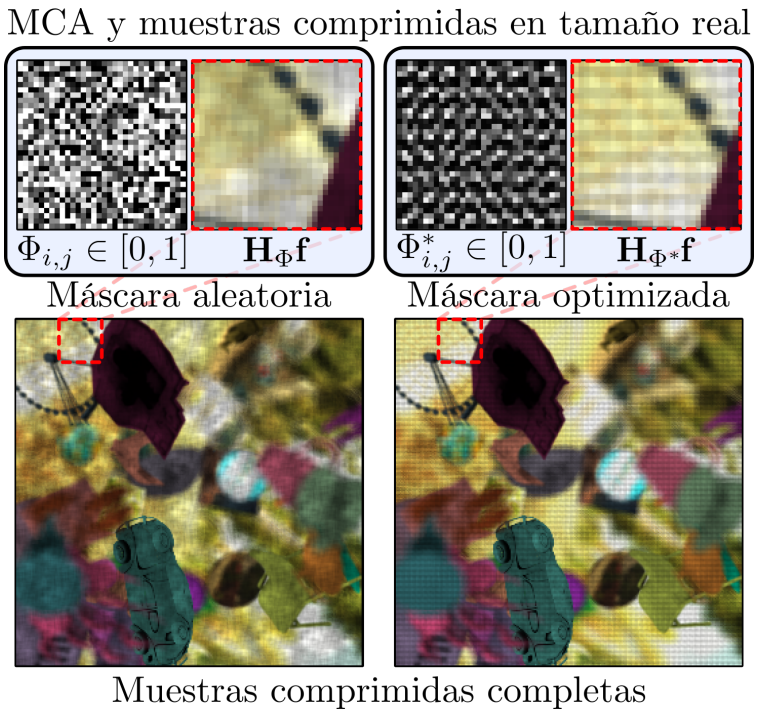
Estimación de los mapas de disparidad.



Nota: De izquierda a derecha se observan pares de columnas (Epinet y Lfattnet) donde se muestran las estimaciones con CLMC y CLAP. Finalmente, se muestran las estimaciones obtenidas con las configuraciones RNC y EAE.

Figura 16

MCAs de tamaño digital real utilizadas en el método propuesto.



Nota: Izquierda: MCA usada para la configuración RNC. Derecha: MCA optimizada, resultado final del entrenamiento de la configuración EAE. Las muestras comprimidas fueron normalizadas por motivos de visualización.

6. Conclusiones

Este trabajo propusó un método para estimar la profundidad a partir de medidas comprimidas de los campos de luz mediante un enfoque EAE, donde se simula el sistema de fotografía de campos de luz como una capa óptica y se estima la profundidad mediante un decodificador. Dos configuraciones para el método propuesto fueron empleadas, un configuración basada completamente en el enfoque EAE y la configuración RNC basada en únicamente entrenar la capa profunda, después de haber obtenido medidas comprimidas de la capa óptica sin entrenar. Experimentalmente, se encontró que el método propuesto EAE es cuantitavemente mejor que la configuración con solo el decodificador (CMA aleatorias) que requiere una mayor complejidad computacional. Adicionalmente, el decodificador propuesto consiste en una U-net amplia con bloques residuales y proyecciones ascendentes. Los resultados obtenidos muestran que el método propuesto, especialmente la configuración EAE, es comparable con algunos de los modelos del estado del arte Shin et al. (2018); Tsai et al. (2020) basados en estimar mapas de disparidad con muestras de campos de luz completos y reconstruidos mediante muestreo compresivo o aprendizaje profundo. Además, el método propuesto es 20 veces más rápido en tiempo de entrenamiento y 23 más rápido en tiempo de inferencia que el modelo Lfattnet entrenado con CLAP, lo que permite concluir que no siempre es necesario realizar reconstrucciones de los datos para realizar la tarea de estimar la profundidad cuando se incorporan los métodos adecuados para el manejo de medidas comprimidas.

Referencias Bibliográficas

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Adelson, E. H., Bergen, J. R., et al. (1991). *The plenoptic function and the elements of early vision*, volume 2. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of .

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322.

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee.

Bacca, J., Galvis, L., and Arguello, H. (2020). Coupled deep learning coded aperture design for compressive image classification. *Optics express*, 28(6):8528–8540.

Bacca, J., Gelvez, T., and Arguello, H. (2021). Deep coded aperture design: An end-to-end approach for computational imaging tasks. *arXiv preprint arXiv:2105.03390*.

Balas, V. E., Roy, S. S., Sharma, D., and Samui, P. (2019). *Handbook of deep learning applications*, volume 136. Springer.

Boyd, S., Parikh, N., and Chu, E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30.

Casser, V., Pirk, S., Mahjourian, R., and Angelova, A. (2019). Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8001–8008.

Chang, J. and Wetzstein, G. (2019). Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10193–10202.

Chen, C., Liu, X., Ding, M., Zheng, J., and Li, J. (2019). 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–192. Springer.

Chen, C.-H. and Chellappa, R. (2017). Face recognition using an outdoor camera network. In *Human Recognition in Unconstrained Environments*, pages 31–54. Elsevier.

Chen, W.-C. (2003). Light field mapping: Efficient representation of surface light fields. *Energy,*

simulation-training, ocean engineering, and instrumentation: research papers of the Link Foundation fellows, page 89.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al. (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10.

De Cubber, G. and Doroftei, D. (2011). Human victim detection and stereo-based terrain traversability analysis for behaviour-based robot navigation. In *Using Robots in Hazardous Environments*, pages 476–498. Elsevier.

Deeba, F., Zhou, Y., Dharejo, F. A., Du, Y., Wang, X., and Kun, S. (2021). Multi-scale single image super-resolution with remote-sensing application using transferred wide residual network. *Wireless Personal Communications*, pages 1–20.

Diamantas, S. C., Oikonomidis, A., and Crowder, R. M. (2010). Depth estimation for autonomous robot navigation: A comparative approach. In *2010 IEEE International Conference on Imaging Systems and Techniques*, pages 426–430. IEEE.

Dong, W., Wang, P., Yin, W., Shi, G., Wu, F., and Lu, X. (2018). Denoising prior driven deep neural network for image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2305–2318.

El Gendy, A., Shalaby, A., Saleh, M., and Flintsch, G. W. (2011). Stereo-vision applications to

reconstruct the 3d texture of pavement surface. *International Journal of Pavement Engineering*, 12(03):263–273.

El Jamiy, F. and Marsh, R. (2019). Survey on depth perception in head mounted displays: distance estimation in virtual reality, augmented reality, and mixed reality. *IET Image Processing*, 13(5):707–712.

Garg, R., Wadhwa, N., Ansari, S., and Barron, J. T. (2019). Learning single camera depth estimation using dual-pixels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7628–7637.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279.

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838.

Gortler, S. J., Grzeszczuk, R., Szeliski, R., and Cohen, M. F. (1996). The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54.

Guan, S., Khan, A. A., Sikdar, S., and Chitnis, P. V. (2019). Fully dense unet for 2-d sparse

photoacoustic tomography artifact removal. *IEEE journal of biomedical and health informatics*, 24(2):568–576.

Guo, M., Hou, J., Jin, J., Chen, J., and Chau, L.-P. (2020). Deep spatial-angular regularization for compressive light field reconstruction over coded apertures. In *European Conference on Computer Vision*, pages 278–294. Springer.

Gupta, M., Jauhari, A., Kulkarni, K., Jayasuriya, S., Molnar, A., and Turaga, P. (2017). Compressive light field reconstructions using deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–20.

Haim, H., Elmalem, S., Giryes, R., Bronstein, A. M., and Marom, E. (2018). Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on Computational Imaging*, 4(3):298–310.

Hajisharif, S., Miandji, E., Guillemot, C., and Unger, J. (2020). Single sensor compressive light field video camera. In *Computer Graphics Forum*, volume 39, pages 463–474. Wiley Online Library.

Harsányi, K., Kiss, A., Majdik, A., and Szirányi, T. (2018). A hybrid cnn approach for single image depth estimation: A case study. In *International Conference on Multimedia and Network Information System*, pages 372–381. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, H., Kuhn, A., Michelini, M., Schmitz, M., and Mayer, H. (2019). 3d urban scene reconstruction and interpretation from multisensor imagery. In *Multimodal Scene Understanding*, pages 307–340. Elsevier.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Inagaki, Y., Kobayashi, Y., Takahashi, K., Fujii, T., and Nagahara, H. (2018). Learning to capture light fields through a coded aperture camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434.

Jain, R., Kasturi, R., and Schunck, B. G. (1995). *Machine vision*, volume 5. McGraw-hill New York.

Kalia, M., Navab, N., and Salcudean, T. (2019). A real-time interactive augmented reality depth estimation technique for surgical robotics. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8291–8297. IEEE.

Kim, H. and Sohn, K. (2003). Hierarchical depth estimation for image synthesis in mixed reality. In

Stereoscopic Displays and Virtual Reality Systems X, volume 5006, pages 544–553. International Society for Optics and Photonics.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE.

Lee, Y., Kim, H., Park, E., Cui, X., and Kim, H. (2017). Wide-residual-inception networks for real-time object detection. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 758–764. IEEE.

Levoy, M. and Hanrahan, P. (1996). Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42.

Li, J. and Jin, X. (2020). Epi-neighborhood distribution based light field depth estimation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2003–2007. IEEE.

Li, L., Fang, Y., Liu, L., Peng, H., Kurths, J., and Yang, Y. (2020a). Overview of compressed sensing: sensing model, reconstruction algorithm, and its applications. *Applied Sciences*, 10(17):5909.

Li, Y., Wang, Q., Zhang, L., and Lafruit, G. (2021a). A lightweight depth estimation network for wide-baseline light fields. *IEEE Transactions on Image Processing*, 30:2288–2300.

Li, Y., Zhang, L., Wang, Q., and Lafruit, G. (2020b). Manet: multi-scale aggregated network for light field depth estimation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1998–2002. IEEE.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021b). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*.

Lin, H., Chen, C., Kang, S. B., and Yu, J. (2015). Depth recovery from light field using focal stack symmetry. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3451–3459.

Liu, X., Chen, P., Du, M., Zang, H., Hu, H., Zhu, Y., Ma, Z., Wang, Q., and Niu, Y. (2020). Multi-information fusion depth estimation of compressed spectral light field images. In *3D Image Acquisition and Display: Technology, Perception and Applications*, pages DW1A–2. Optical Society of America.

Liu, X., Wang, Q., Ma, Z., Niu, Y., Duan, S., Zang, H., Ma, F., Huang, M., Lv, Q., and Liang, E. (2018). Depth estimation and multi-view spectral images based on compressive sensing light field reconstruction. In *3D Image Acquisition and Display: Technology, Perception and Applications*, pages 3Tu3E–5. Optical Society of America.

Ma, H., Li, H., Qian, Z., Shi, S., and Mu, T. (2018). Vommanet: An end-to-end network for

disparity estimation from reflective and texture-less light field images. *arXiv preprint arXiv:1811.07124*.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mahjourian, R., Wicke, M., and Angelova, A. (2018). Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675.

Marwah, K., Wetzstein, G., Bando, Y., and Raskar, R. (2013). Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):1–12.

Matlab, S. (2012). Matlab. *The MathWorks, Natick, MA*.

Mun, J.-H. and Ho, Y.-S. (2018). Depth estimation from light field images via convolutional residual network. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1495–1498. IEEE.

Nabati, O., Mendlovic, D., and Giryes, R. (2018). Fast and accurate reconstruction of compressed color light field. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE.

Nakayama, Y., Huimin, L., Yujie, L., and Hyoungseop, K. (2018). Wide residual networks for semantic segmentation. In *2018 18th International Conference on Control, Automation and Systems (ICCAS)*, pages 1476–1480. IEEE.

Nalpantidis, L. and Gasteratos, A. (2012). Stereo vision depth estimation methods for robotic applications. In *Depth Map and 3D Imaging Applications: Algorithms and Technologies*, pages 397–417. IGI global.

Nam, K. W., Park, J., Kim, I. Y., and Kim, K. G. (2012). Application of stereo-imaging technology to medical field. *Healthcare informatics research*, 18(3):158.

Ni, Z.-L., Bian, G.-B., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Wang, C., Zhou, Y.-J., Li, R.-Q., and Li, Z. (2019). Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In *International Conference on Neural Information Processing*, pages 139–149. Springer.

OMahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., and Walsh, J. (2019). Deep learning vs. traditional computer vision. In *Science and Information Conference*, pages 128–144. Springer.

Pope, G. (2009). Compressive sensing: A summary of reconstruction algorithms. Master's thesis, ETH, Swiss Federal Institute of Technology Zurich, Department of Computer .

Qaisar, S., Bilal, R. M., Iqbal, W., Naureen, M., and Lee, S. (2013). Compressive sensing: From theory to applications, a survey. *Journal of Communications and networks*, 15(5):443–456.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404.

Rogge, S., Schiopu, I., and Munteanu, A. (2020). Depth estimation for light-field images using stereo matching and convolutional neural networks. *Sensors*, 20(21):6188.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Sanz, P. R., Mezcua, B. R., and Pena, J. M. S. (2012). Depth estimation-an introduction. In *Current Advancements in Stereo Vision*. IntechOpen.

Sapijaszko, G. and Mikhael, W. B. (2018). An overview of recent convolutional neural network algorithms for image recognition. In *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 743–746. IEEE.

Sawano, Y., Miura, J., and Shirai, Y. (2001). Man chasing robot by an environment recognition using stereo vision. In *Human Friendly Mechatronics*, pages 241–246. Elsevier.

Schambach, M. and Heizmann, M. (2020). A multispectral light field dataset and framework for light field deep learning. *IEEE access*, 8:193492–193502.

Shedligeri, P. A., Mohan, S., and Mitra, K. (2017). Data driven coded aperture design for depth

recovery. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 56–60. IEEE.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2021). Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539.

Sheng, H., Zhao, P., Zhang, S., Zhang, J., and Yang, D. (2018). Occlusion-aware depth estimation for light field using multi-orientation epis. *Pattern Recognition*, 74:587–599.

Shin, C., Jeon, H.-G., Yoon, Y., Kweon, I. S., and Kim, S. J. (2018). Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4748–4757.

Smolyanskiy, N., Kamenev, A., and Birchfield, S. (2018). On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1007–1015.

Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., and Bouaziz, S. (2021). Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372.

Tsai, Y.-J., Liu, Y.-L., Ouhyoung, M., and Chuang, Y.-Y. (2020). Attention-based view selection

networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12095–12103.

Vadathya, A. K., Cholleti, S., Ramajayam, G., Kanchana, V., and Mitra, K. (2017). Learning light field reconstruction from a single coded image. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 328–333. IEEE.

Vadathya, A. K., Girish, S., and Mitra, K. (2019). A unified learning-based framework for light field reconstruction from coded projections. *IEEE Transactions on Computational Imaging*, 6:304–316.

Valloli, V. K. and Mehta, K. (2019). W-net: Reinforced u-net for density map estimation. *arXiv preprint arXiv:1903.11249*.

Wang, L., Zhang, T., Fu, Y., and Huang, H. (2018). Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270.

Wang, Y., Lai, Z., Huang, G., Wang, B. H., Van Der Maaten, L., Campbell, M., and Weinberger, K. Q. (2019). Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900. IEEE.

Wu, G., Masia, B., Jarabo, A., Zhang, Y., Wang, L., Dai, Q., Chai, T., and Liu, Y. (2017). Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954.

Wu, L., Jaiprakash, A., Pandey, A. K., Fontanarosa, D., Jonmohamadi, Y., Antico, M., Strydom, M., Razjigaev, A., Sasazawa, F., Roberts, J., et al. (2020). Robotic and image-guided knee arthroscopy. In *Handbook of Robotic and Image-Guided Surgery*, pages 493–514. Elsevier.

Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., and Veeraraghavan, A. (2019). Phase-cam3dlearning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE.

Xu, D., Liu, X., and Zhang, Y. (2020). Real-time depth estimation for aerial panoramas in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 705–706. IEEE.

Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., and Yang, G.-Z. (2017). Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*.

You, Y., Wang, Y., Chao, W.-L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., and Weinberger, K. Q. (2019). Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, Q., Yang, L. T., Chen, Z., and Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, 42:146–157.

Zhang, Z. (1998). Determining the epipolar geometry and its uncertainty: A review. *International journal of computer vision*, 27(2):161–195.

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858.

Zhu, H., Wang, Q., and Yu, J. (2017). Occlusion-model guided antiocclusion depth estimation in light field. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):965–978.

Zhu, K., Xue, Y., Fu, Q., Kang, S. B., Chen, X., and Yu, J. (2018). Hyperspectral light field stereo matching. *IEEE transactions on pattern analysis and machine intelligence*, 41(5):1131–1143.