End-to-End Optimization of a Coded Stereo Imaging System for Depth Estimation

Jhon Edinson López Durán

Trabajo de Grado para optar al título de Magíster en Ingeniería de Sistemas e Informática

Director

Henry Arguello Fuentes

Doctorado en Ingeniería Eléctrica y Computación

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Bucaramanga

2023

Dedicatoria

Este trabajo viene dedicado a mi madre Luz Yanneth, a mi padre Pedro Jesús y mi hermano Jesús Eduardo por su apoyo incondicional.

Agradecimientos

Agradezco a mi familia por el apoyo ecónomico y moral que tuvieron para conmigo durante el desarrollo de mi carrera. También agradezco a mis amigos y compañeros por las vivencias de estos inolvidables años de universidad.

Un reconocimiento y agradecimiento importante lo realizo a mi compañero de posgrado Edwin, a mi director de trabajo de grado Henry, por dedicar su tiempo, experiencia y conocimiento en la guía de mi proyecto.

Tabla de Contenido

List of Figures

List of Tables

Resumen

Título: Optimización de Extremo a Extremo de un Sistema Estéreo de Adquisición de Imágenes Codificadas para la Estimación de la Profundidad [*]

Autor: Jhon Edinson López Durán [**]

Palabras Clave: Estimación de la Profundidad, Optimización de Extremo a Extremo, Visión Estereoscópica.

Descripción: La estimación de la profundidad es esencial para la comprensión de escenas, la conducción autónoma, la robótica y otras áreas. Sin embargo, la estimación de la profundidad sigue siendo un reto debido a la pérdida de información 3D durante el proceso de captura de imágenes RGB. A lo largo de los años, se han propuesto diferentes arquitecturas ópticas para capturar las escenas, desde una sola cámara hasta múltiples cámaras con iluminación activa o pasiva. También hay que desarrollar arquitecturas y métodos para la estimación de la profundidad en cualquier entorno. En este sentido, la adquisición de imágenes estereoscópicas es una arquitectura de adquisición eficiente, ya que imita el sistema de visión humano y estima la profundidad mediante estereopsis. Aunque las redes neuronales profundas mejoran el rendimiento de la estimación de la profundidad, sigue habiendo dificultades para predecir la profundidad absoluta y generalizar fuera de un entorno predeterminado. Por ello, recientemente se ha propuesto un enfoque denominado óptica profunda, que diseña elementos ópticos, como máscaras de fase y lentes difractivas, junto con el algoritmo de procesamiento de imágenes de forma integral. Por lo tanto, este trabajo propone utilizar el paradigma de la óptica profunda en la estimación de profundidad estereoscópica mediante el diseño de aperturas codificadas bajo un enfoque de optimización de extremo a extremo. El enfoque profundo propuesto se evalúa utilizando un conjunto de datos de última generación y, además, el método propuesto se valida en una configuración óptica real.

---

[*] Trabajo de Maestría

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Henry Arguello Fuentes, Doctorado en Ingeniería Eléctrica y Computación.

Abstract

Title: End-to-End Optimization of a Coded Stereo Imaging System for Depth Estimation [*]

Author: Jhon Edinson López Durán [**]

Keywords: Depth Estimation, End-to-End Optimization, Stereoscopic Vision.

Description: Depth estimation is essential for scene understanding, autonomous driving, robotics, and other areas. However, depth estimation remains challenging because of losing the 3D information during the RGB image capture process. Over the years, different optical architectures have been proposed to capture the scenes, from a single camera to multiple cameras with active or passive illumination. Also, architectures and methods for depth estimation in any environment have to be developed. Accordingly, stereoscopic image acquisition is an efficient acquisition architecture, as it mimics the human vision system and estimates depth by stereopsis. Although deep neural networks improve depth estimation performance, there are still difficulties in predicting absolute depth and generalizing outside a predetermined environment. Therefore, an approach called deep optics has recently been proposed, which designs optical elements, such as phase masks and diffractive lenses, in conjunction with the image processing algorithm in an end-to-end manner. Therefore, this work proposes using the deep optics paradigm in stereo depth estimation by designing coded apertures under an end-to-end optimization approach. The proposed deep approach is evaluated using a state-of-the-art data set, and additionally, the proposed method is validated in a real optical setup.

---

[*] Master Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Henry Arguello Fuentes, Doctorado en Ingeniería Eléctrica y Computación.

Introduction

Vision is one of the five human senses and arguably the most important of them. This sense allows us to obtain large amounts of information about our environment, which are essential for our survival. Thanks to it, we can identify objects and animals within our field of vision at any given moment and thus estimate their position so that our brain can assess their usefulness and importance. However, the vision is also one of the most complicated senses to reproduce artificially because cognitive psychology has not fully understood it Dornaika (1995). Computer vision (CV) is an interdisciplinary scientific field that deals with how computers can obtain a high-level understanding of an environment through images or digital videos. From the engineering perspective, CV seeks to understand and automate some tasks related to the human visual systemHuang (1996); Sonka et al. (2014). Computer vision tasks include acquiring, processing, analyzing, and understanding digital images and extracting high-dimensional data from the real world to produce numerical or symbolic information, for example, decision forms Morris (2004); Klette (2014), depth estimation Haim et al. (2018), image segmentation Haralick and Shapiro (1985) among others Umbaugh (2010). Understanding in this context means transforming visual images (the entry of the retina) into descriptions of the world that make sense of thought processes and can cause appropriate action. This understanding of the image can be seen as the debauchery of the symbolic information of the image data using models built with geometry, physics, statistics, and learning theory Forsyth and Ponce (2002). For instance, depth estimation is a task that consists of identifying the depth of different objects present in the scene with respect to the camera.

Digital images represent the 3D real world in two spatial dimensions (x, y); this means that between the passage from the natural scene to the captured image, information about the depth z of the objects in the scene is lost. The stereoscopic vision, a sub-branch of computer vision, constitutes an alternative for obtaining the lost depth information. The motivation of this stereoscopic approach is to imitate the human visual system that can perceive its surroundings in three dimensions and is constituted by two eyes; this has made artificial stereoscopic systems use at least two different cameras to acquire two images of the same scene. It is possible to determine the distance at which any object contained in the two images is located with respect to the observer. Cameras are used to capture images, and algorithms are required to perform the calculations that determine the distance to the observer. Focusing on artificial systems, the information about the distances at which the objects in the scene are located is generated in the form of a structure known technically as a disparity map. Subsequently, using a simple geometric relation by triangle similarity and knowing some parameters of the cameras, such as the existing separation between them (baseline) and the focal distances of their optical systems, it is possible to determine the object depths. Obtaining the disparity map requires identifying the same point or object in the two images, representing the same physical entity in the three-dimensional scene. The process by which the same three-dimensional entity is identified in both images is known as stereoscopic correspondence. In all processes of stereoscopic vision, the correspondence constitutes the current problem to which a very high percentage of research has been dedicated in the field of stereoscopic vision.

Some methods for this task include area-based measures, which find a correlation between image patches Hannah (1988). However, one of the most effective methods is to analyze each pixel as a group of collected data points that an algorithm can identify. This allows an analysis of each pixel from all cameras. This requires a large amount of processing power. Other approaches are line matching on boundaries or edges of the object being captured, matching all common features between images (i.e., 100 % of pixels that have captured the same features are matched) Oâ€™Riordan et al. (2018). Finally, the coincidence of characteristics is the fundamental base adopted in recent years through the use of convolutional neuronal networks (CNNs), which learn to extract the features of the images and seek their corresponding pair between them Song et al. (2018). Recent works have also used CNNs to explore depth cues using the features learned by CNNs, improving approaches based on hand-crafted features Eigen et al. (2014); Fu et al. (2018). In contrast, it has also been shown that coded defocus blur or chromatic aberrations can generate powerful depth cues and help estimate an accurate depth map. Thus, imaging systems that employ an amplitude or phase coding mask have been proposed in the last two decades to promote depth cues obtained using an optical system with a depth-dependent point spread function (PSF) Haim et al. (2018); Levin et al. (2007); Chang and Wetzstein (2019). However, these systems employ specialized optical encoding strategies sensitive to ambient light, limiting outdoor usage.

On the other hand, conventional encoded-depth imaging systems design the optical system first. Then the image processing algorithm parameters are tuned or learned to obtain the best depth map recovery. Nevertheless, a more recent framework dubbed deep optics introduces the

concept of optimizing the optical elements jointly with the reconstruction algorithm in an end-to-end fashion employing stochastic optimization. Based on differentiable reconstruction algorithms, this framework allows the optimization of domain-specific optical elements. This idea has been recently investigated in applications such as extended depth of field Sitzmann et al. (2018), high dynamic range imaging Metzler et al. (2020), depth estimation Chang and Wetzstein (2019), and many others. Most methods in which depth is encoded and the optical elements are designed using end-to-end optimization have been developed for monocular systems. This system calculated the depth map from a single RGB image Bhoi (2019).

The depth encoding information on stereo systems has also been studied previously. For instance, in Wang et al. (2014), the authors explore the improvement achieved in depth estimation by merging coded apertures in a stereo system using heuristic algorithms. This is analyzed using several stereo camera setups equipped with different coded apertures to infer such possibilities. The employed coded apertures are selected randomly and independent of the algorithm used to estimate the depth information. These analysis results are encouraging because coded apertures can provide valuable complementary information to stereo-based depth estimation. In particular, utilizing the inherent relation between stereo cues and defocus cues extracts depth information more robustly, especially for problematic scene regionsWang et al. (2014).

On the other hand, using the end-to-end optimization approach, Shiyu et al. Tan et al. (2021) propose a CNN to estimate the disparity maps and jointly design a phase mask to increase the

depth of field in a stereo vision system, and with this, obtain a more accurate disparity map for far distances, achieving through computational simulations and real hardware implementation a high accuracy. Although these two approaches mentioned above propose a new methodology for depth estimation, these approaches have a big difference, reflected in the implementation and calibration costs. Therefore, in this research work, we propose to merge the best of these two approaches, thus proposing a new deep learning architecture for disparity estimation, which does not require high computational power compared to traditional CNNs, and using this proposed architecture under an end-to-end optimization approach, design a pair of coded apertures, which promote depth-cues, thus achieving a new low-cost and accurate strategy for depth estimation in a stereo vision system.

1. Objectives

General Objective

To design a coded aperture computational stereo imaging system for depth estimation by using an end-to-end optimization of the sensing and the depth map recovery processes.

Specific Objectives

To develop the optical model that describes the light propagation process, from the source to the camera sensor, passing through the optical elements.

To implement a neural network for the estimation of the depth from the coded stereo images.

To learn the coded aperture of the proposed stereo imaging system jointly with the parameters of the convolutional neuronal network.

To validate the obtained coded stereo system employing computational simulations and real implementation in the optics laboratory.

2. Theoretical Background

This section describes some applications in which it is necessary to know the depth of the objects in a scene, followed by the general concepts of an image formation system. The pillars of a stereo vision system are shown, and the geometry of a stereo system with its parallel axes is known as canonical configuration for depth estimation. Then the different approaches that have existed over the years for estimating disparity maps are shown, and finally, the concept of encoded depth imaging is introduced.

2.1. Depth Maps Applications

2.1.1. 3D Tracking. 3D tracking uses the depth maps obtained to refine and enhance the scene information obtained from the images. This technique can provide information on specific objects in the image, identifying them separately within the map (Haritaoglu et al., 1998). Such information is used to know the trajectory and position of these objects when there are indications that they are moving. This requires several depth maps at different points in time. 3D tracking is used for 3D reconstruction of objects, security and surveillance applications where it is necessary to know the spatial positions of individuals in the scene, and many other applications.

2.1.2. SLAM and SFM. Simultaneous Localization and Mapping (SLAM) aims at the simultaneous resolution of the location and mapping of elements in the environment, i.e., the resolution of the three-dimensional features of a scene, and providing information about the positions from which the images are taken (Grisetti et al., 2010). SLAM attempts to provide a three-dimensional structure in real-time using information from different sensors or systems, with

Figura 1. Tracking with orientation snapping (Chang et al., 2019). Using lane direction information helps to determine the vehicle orientation for detection and tracking. Ground truth cuboids are green.

depth maps being one input system. There is a visual variant of SLAM (vSLAM), which uses only images to solve the 3D reconstruction and obtain the position. This method is an application of the Structure from Motion (SFM) technique (Ullman, 1979). However, the latter can offer a much more complete resolution than SLAM, although more challenging to implement in real time.

2.1.3. Autonomous Driving. Automated driving in automobiles is a topic of great interest today, as vehicles that can guide themselves and move on the road autonomously serve as excellent support for drivers, for example, on long journeys. Autonomous driving encompasses various systems that can be implemented, from small robots to large, but all motorized, vehicles. Thus, autonomous driving creates and maintains a map of its environment based on a series of sensors located in different parts of the device. Radar sensors monitor the position of nearby vehicles (Boric et al., 2021). Video cameras detect traffic lights, read traffic signs, follow other vehicles and

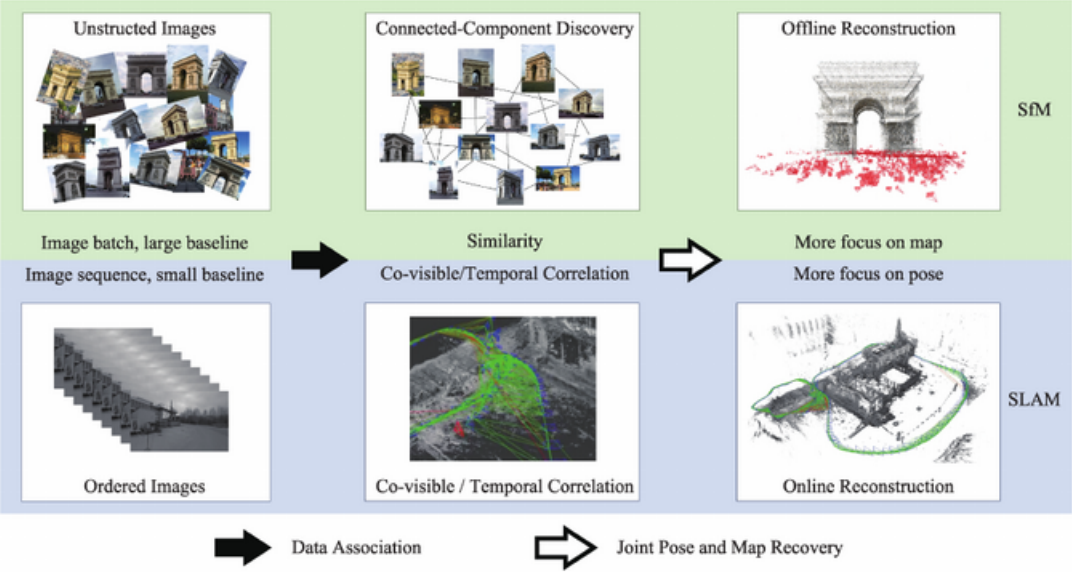

Figura 2. Similar pipeline of SFM and SLAM (Yan and Zha, 2019).

look for pedestrians. The systems of depth measure distance to detect road edges and identify lane markings and the proximity of different road actors.

Figura 3. Performance of the different sensors for driving an autonomous vehicle [1].

[1] Image obtained from `https://researchleap.com/`.

The applications mentioned above are just some of the applications that depth maps have, as these maps have applications in other areas such as medicine, agriculture, virtual reality, fog and haze removal, and Parallax Simulation, among others. Once the variety of applications and the importance of depth maps are shown, the basic concepts necessary to estimate depth maps using a stereo vision system will be shown.

2.2. Single Camera

To acquire an image of space, it is necessary that the light beams reflected on the objects pass through a small hole so that each point in space projects a single beam of light to the other side, projecting itself onto the so-called image plane. Fig. 4 illustrates this process. The optics involved leads to treating the image inverted; however, in practice, it is simpler to work with a mathematically equivalent system, placing the image plane in front of the center of projection (see Fig. 5). This new configuration respects the same original geometrical properties (Calderón Saavedra, 2012).

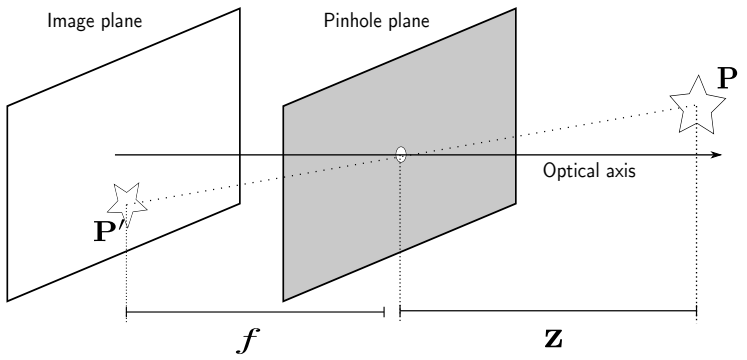

Figura 4. Model pinhole, where **z** is the depth of the point **P**. **f** is the focal length, and $\mathbf{P}'$ is the projection of the point **P** in the image plane.

In Fig. 5, some concepts, such as the projection center, which corresponds to the reference

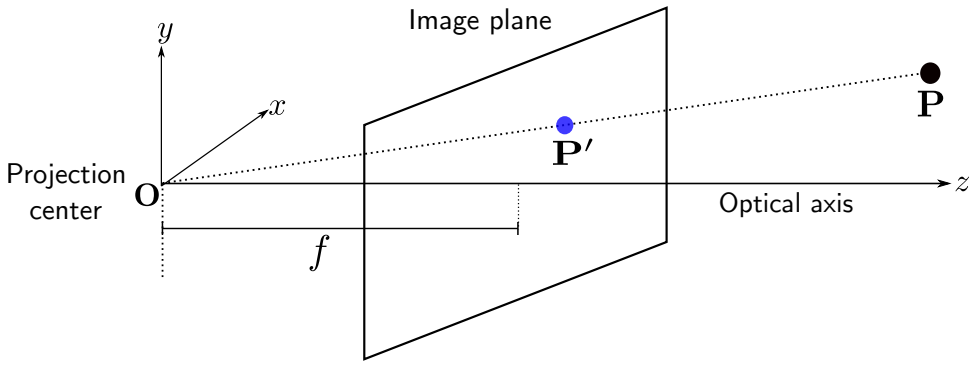

Figura 5. Representation of a traditional coordinate system in an image acquisition system.

system associated with the camera; the plane image in which the observed scene is projected; finally, the optical axis or central ray, which is born in the center of the projection and intercept perpendicularly on the image plane. Notice that the optical axis goes through the image (in pixel units) just in the center; also, the image belongs to the image plane, but it works in coordinates of pixels and non-metrics (Calderón Saavedra, 2012).

Analyzing the previous figure, three reference systems can be generated, which are:

- **Camera:** with origin in the projection center and metric units. The axis, the image plane in the position of the optical center. Axes *x* e *y* go through the horizontal and vertical directions, respectively. This system works on a plane $z = f$ that is considered two-dimensional. $\mathbf{P}'$ is used to represent a point in this system.

- **Image:** it corresponds to a transformation of the system associated with the camera in such a way as to work in pixel units. Its origin is located in the upper left corner of the image, indexing *i* to the right and *j* down. $\mathbf{q}$ is used to represent a point in this system.

- **Scene:** this is the global reference system used; the real space points belong, therefore, work

in metric units. Usually, its origin coincided with the camera projection center. The x, y, z variables are used to describe this space. **P** is used to represent a point in this system.

The imaging model illustrated above is known as a pinhole camera. This model assumes that a single light beam can pass through such a barrier. However, if the aperture is too small, not enough light is gathered in a brief exposure, and diffraction phenomena may arise; if the aperture is too large, the image becomes blurred since a single point in space is projected multiple times due to many light beams passing through the barrier. Therefore, more than a pinhole is needed to capture an image with an excellent spatial resolution; this problem is solved using a lens that concentrates an appropriate amount of light on the image plane.

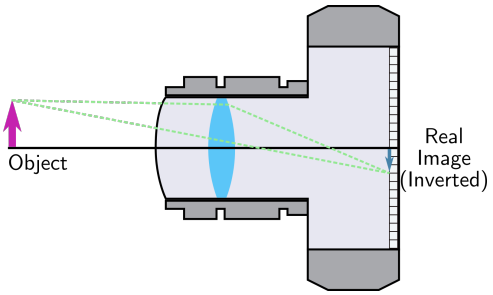

Figura 6. Conventional image formation model using convex lens and ray tracing.

2.3. Stereo Vision

Binocular vision is defined as the vision from two eyes where some amount overlaps the data being perceived from each. The overlap between the two different views is used in biological vision to perceive depth. Stereoscopic vision is the use of binocular vision to perceive the three-dimensional structure of the world. Binocular disparity is the difference in the placement of objects as viewed by two eyes due to the different viewpoints from which each views the world. Stereopsis

is the impression of depth extracted from binocular disparity (Howard et al., 1995). A stereo vision system is a set of two or more cameras used by machines to extract the depth of a 3-D scene as viewed from different vantage points, as modeled after binocular vision in humans. Fig 7 shows how stereo vision is present in humans by using two eyes viewing a scene from different vantage points to extract depth. In humans, this is known as depth perception, where the human brain merges the left-eye view and the right-eye view to form a single image.

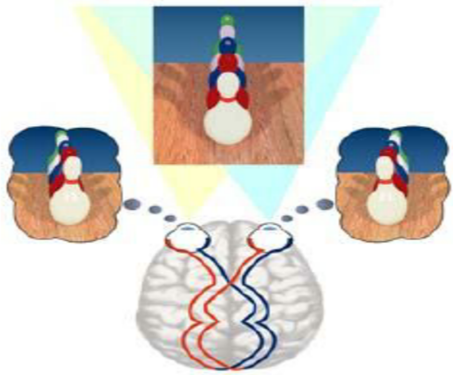

Figura 7. Example of binocular vision in humans (Short, 2009).

2.3.1. Setup of Stereo System. In a stereo vision system, the cameras are aligned horizontally and separated by a distance known as **baseline**. Fig. 8 shows an example of a stereo vision system with two cameras mounted on a bar. These two cameras provide the two images needed to extract a disparity map, which provides the necessary data to estimate the depth map of the scene.

Once we obtain the set of images using the optical configuration shown in Fig. 8, these

---

[2] Image obtained from `https://nerian.com/products/stereo-vision-accessories/`.

Figura 8. Real stereo vision system with two cameras[2].

images are the input to an algorithm to estimate their disparity map. Next, in Fig. 9, we illustrate by triangulation how we can estimate the depth of a point **P** at a distance z from the cameras by estimating its disparity.

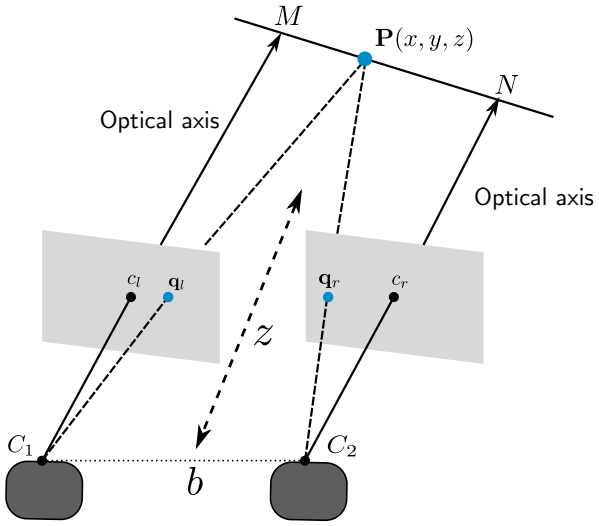

Figura 9. The stereo vision geometry, where the points *M* and *N* represent the intersection of the optical axes for each camera with a parallel line to it. The points c_l and c_r represent the intersection de optical axes with the center of the virtual image plane, and C_r and C_l represent the central points in the cameras. Subindex l and r denote left and right.

Both cameras observe the same point **P** in space, projected as q_l and q_r in their respective virtual image plane. Knowing the parameters of the stereo system, such as the focal length f and

the baseline b, employing triangulation, we can estimate at what depth z is the point **P** with respect to cameras as we can see from the diagram in Fig. 9, the camera plane is parallel to the virtual image plane, then we have:

$$\Delta C_1 MP \sim \Delta C_1 q_l c_l, \tag{1}$$

and for the rigth camera we have:

$$\Delta C_2 NP \sim \Delta C_2 q_r c_r. \tag{2}$$

From Eq. 1, we know that:

$$\frac{z}{f} = \frac{MP}{c_l q_l}, \tag{3}$$

and from Eq. 2:

$$\frac{z}{f} = \frac{NP}{c_r q_r}. \tag{4}$$

Since baseline, b is the distance between the two cameras in a stereo unit,

$$\therefore b = MP - NP. \tag{5}$$

From Eqs. 3 and 4 we can rewrite the Eq. 5 as:

$$b = \frac{z}{f}(c_l q_l - c_r q_r), \tag{6}$$

where $(c_l q_l - c_r q_r)$ is nothing but disparity d. Therefore, from Eq. 6 we arrive at the original equation of depth, i.e.:

$$z = \frac{f \times b}{d}. \tag{7}$$

The proof for the above equation implies that the depth from the stereo unit is only dependent on the stereo focal length, the baseline length, and the disparity between the corresponding pixels in the image pair. For this reason, stereo depth estimation is more robust and better suited. It is independent of any orientation or poses of the stereo unit to the scene in the 3D world coordinates. The depth of an object shown by the stereo unit is not affected by any movement of the unit at the same distance from the object. This characteristic does not hold when the depth is estimated using the monocular camera methods other than deep learning (Praveen, 2020).

Calculating depth using a monocular camera is highly dependent on the exact pose of the unit to the scene in the 3D world coordinates. The pose constants that work for depth estimation in one pose of the camera are most certainly guaranteed not to work when the camera is repositioned to some other pose at the same depth from the object. Additionally, Eq. 7 highlights the inverse proportionality between z and d. In this way, objects close to cameras produce significant disparities. They can be well differentiated, while in the distance, they have a slight disparity associated

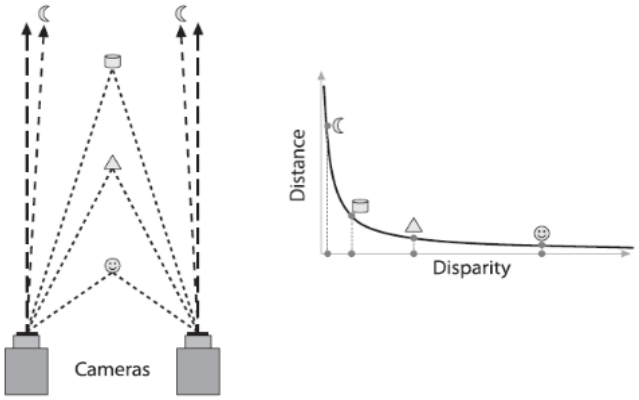

Figura 10. Distance and disparity are inversely proportional (Bradski and Kaehler, 2008).

with and can inevitably be interpreted as a plane.

2.4. Algorithms for disparity estimation

As seen in Eq. 7 and Fig. 10, the depth at which an object is located is related to the disparity. Therefore, all the depth estimation algorithms in stereo systems focus on disparity estimation. As a final step, Eq. 7 is applied to estimate the depth map of the captured scene. Below are some state-of-the-art methods which are used to estimate the depth.

2.4.1. Image Matching. Many approaches have been applied to image matching to optimize the task; however, some have proven more successful than others. Some methods include global optimization, algorithms based on dynamic programming, or 2-D curve matching (Hannah, 1988). Area-based measures treat one image as a reference image. Within this image, a window will be analyzed and statistically compared to other images captured simultaneously from other cameras, often referred to as target images. Image hierarchy, such as the coarse-fine approach, may be utilized at this point to obtain an accurate match and then allow for precise image capture. Usually, a difference-based metric will be used, such as a minimized RMS or maximized correla-

tion (e.g., mean and variance normalized cross-correlation) (Hannah, 1988; Barnard and Fischler, 1982). While more modern approaches provide many real-time stereo systems, most run on a correlation stereo engine (Gehrig et al., 2009). Correlation-based methods allow for high confidence in the procedure but diminish reliability when longer baselines are used, image texture is problematic, or high noise occurs. Global optimization techniques employ an image hierarchy. Global optimization will match all points within the two images that correspond (Barnard, 1987).

2.4.2. False Boundary, Boundary Overreach & Reflectivity Issues. While an adaptive window, a multiple window/Symmetric window, and a shrinking window have been applied to solve false boundary and boundary overreach problems, they have proven to harm smooth object surfaces within the acquired 3D image, resulting in a trade-off between the smoothness of surfaces and the precision of a boundary (Okutomi et al., 2002).

Adaptive window techniques address the problem of having a window large enough to include intensity variation for matching purposes and small enough to prevent projective distortion. Kanade et al. (Kanade and Okutomi, 1994) developed an adaptive window technique by evaluating the local intensity and disparity variation, which allows a suitably sized window to minimize the uncertainty of disparity within a region. Multiple window or symmetry window techniques perform multiple analyses on windows surrounding a pixel area, aiming to find the window with the lowest sum of squared difference (SSD). This window is more likely to cover a constant depth; matching is carried out at a base point within this window (Fusiello et al., 1997).

2.4.3. Obstacle Identification. Stereo vision has become more common as a solution for robotic navigation; however, obstacle identification has presented a challenge, as many obstacles, such as tabletops, can be identified incorrectly as shadows and shadows as obstacles. Murray et al. (Murray and Little, 2000) identified this problem and addressed it using the following logic. If a spike is locally stable but insignificant and lacks support from surrounding surfaces, it is treated as a spiking error and eliminated from the analysis process. If the spike shows a lack of disparity discontinuities at all corners, is locally consistent, and shows visual evidence of being globally part of a larger 3D surface, it is a solid obstacle. Segmentation of images allowed for a hypothesis to be formed on the size of a continuous 3D object (i.e., connected pixels which show local consistency) (Oâ€™Riordan et al., 2018). This rejects false noise spikes that may pass noise filters, but also it allows for the recognition of thin structures which belong to a part of a larger coherent structure (Murray and Little, 2000).

2.4.4. Deep Learning. All the methods discussed above are ultimately static methods that work on traditional computer vision. Deep learning, gaining popularity in recent years, has shown promising results in almost all fields that it has been applied. Sticking to the trend, the researchers and experts also used it to estimate depths and disparity. As expected, the results are encouraging enough for all enthusiasts for further motivated research (Praveen, 2020). Therefore, in depth estimation from a stereo image pair, there are two CNN architectures to perform this task: one is an encoder-decoder architecture, which estimates the disparity as a pixel-level classification problem, and the other approach uses the regularization of a cost volume over a CNN to minimi-

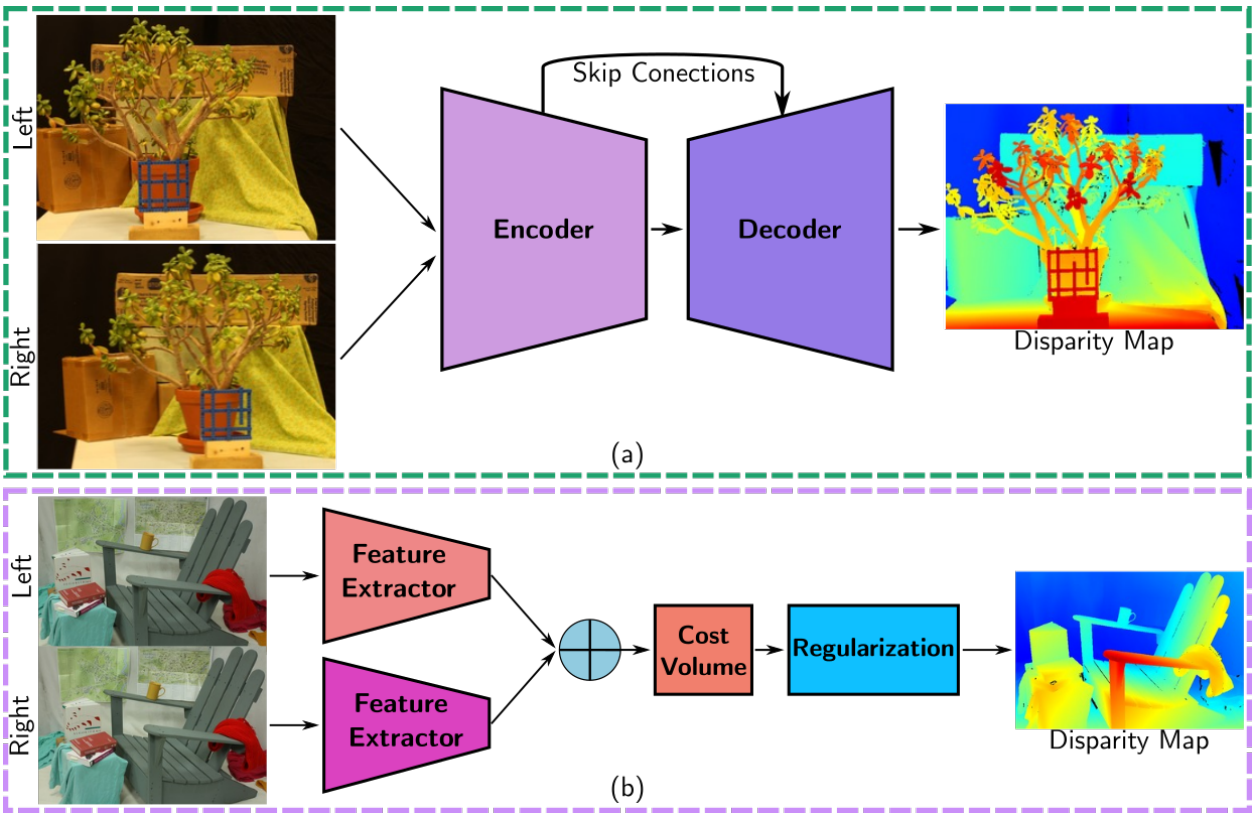

Figura 11. Convolutional neuronal network architectures for disparity estimation. (a) The model learns to map stereo pairs to the corresponding disparity map. (b) Fusing features from the left and right images to regularize to yield the optimal disparity map.

ze the difference between the ground truth and estimated disparity maps. Additionally, Kun et al. (Zhou et al., 2020) classify neural networks into two categories:

- **Non-end-to-end learning:** This approach introduces convolutional neural networks to substitute one or more components in the legacy stereo pipeline. Zbontar and Le Cun (Zbontar and LeCun, 2015) first successfully substituted handcrafted matching cost metrics with deep metrics and achieved considerable gain compared to traditional approaches in terms of both accuracy and speed. They introduced a deep-Siamese network to measure the similarity between two 9×9 image patches. Later, Luo et al. (Luo et al., 2016) accelerated matching

cost calculation by introducing an inner-product layer and treated the patch matching as a multilabel classification problem.

These CNNs and other existing models have achieved significant profits compared to traditional methods. However, the limitations of stereo networks are the following: high computational burden from multiple forward passes for all potential disparities. The narrow receptive field and the lack of context information to infer reliable correspondences in ill-posed and still using post-processing functions are hand-engineered with several empirically set parameters.

- **End-to-end learning:** with the success of Mayer et al. (Mayer et al., 2016), end-to-end stereo-matching networks have become increasingly popular in stereo-matching algorithms. With this approach of convolutional neural networks designing and supervising the network, a refined disparity could also be obtained without post-processing. Many algorithms based on this have been proposed and generate highly accurate depth estimates from stereo image pairs. These methods could roughly be categorized into two groups: 2D encode-decoder structures and regularization modules composed of 3D convolutions (Zhou et al., 2020).

  However, state-of-the-art stereo methods still need help finding correct correspondences in textureless regions, detailed structures, small objects, and near boundaries. Moreover, end-to-end stereo matching net-works-based approaches require colossal memory and high datasets with corresponding ground truth disparity-depth data for training, which means massive data labeling work.

2.5. Encoded Depth Imaging

Optical encoding has allowed the improvement of depth estimation. These optical encodings can be generated by designing coded aperture (CA), phase masks, and diffractive lenses designed to promote depth cues as coded defocus-blur or chromatic aberrations, thus generating an optical system that responds differently depending on the depth of the scene (Chang and Wetzstein, 2019).

2.5.1. Coded-Aperture Based Imaging. Coded aperture imaging (CAI) is a two-stage imaging process. The coded image is obtained by convolving the object from the point spread function (PSF) of the intensity of the coded aperture. This image captured by the sensor is not the desired final image, but is encoded to facilitate information extraction. More specifically, the blur produced by a lens can be controlled using coded apertures that allow both extractions of depth information and retrieval of a standard image, Levin et al. (Levin et al., 2007).

Figure 12 shows a simplified thin lens model that projects the scene light rays onto the sensor to help the viewer comprehend how blur can be managed and utilized. All the rays from a point in the scene will converge to a single sensor point when an item is positioned at the focus distance D, producing a sharp output image. A blurred image is produced when rays from an object at a distance D that is beyond the focus distance strike several sensor sites. The circle of confusion is a common name for the pattern of this blur, which is determined by the aperture cross-section of the lens. The amount of defocus, characterized by the blur radius, depends on the distance of the

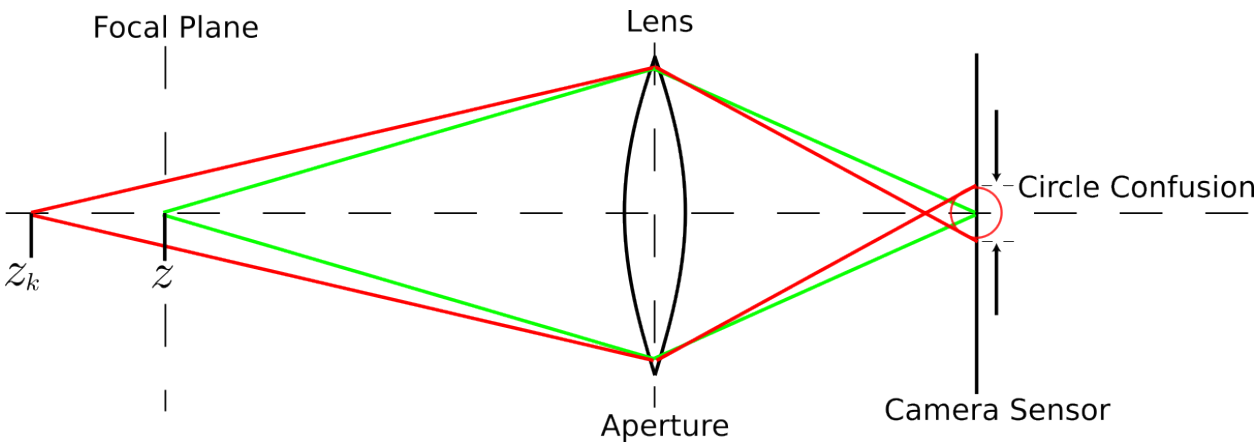

Figura 12. A 2D thin lens model. At the plane of focus, a distance Z from the lens, light rays (shown in green) emanating from a point are focused to a point on the camera sensor. Rays from a point at a distance Z_k (shown in red) no longer map to a point, but rather to a region of the sensor, known as the circle of confusion. The pattern within this circle is determined by the aperture shape (Levin et al., 2007).

object from the focus plane.

For a simple planar object at a distance D, the imaging process can be modeled as a convolution:

$$y = k * x, \tag{8}$$

where y is the observed image, x is the true sharp image, and the blur filter k is a scaled version of the aperture shape. Fig. 13(a) shows the pattern of blur from a conventional lens, the pentagonal disk shape being formed by the intersecting diaphragm blades. The defocus from such aperture does provide depth cues, e.g., (Pentland, 1987). Still, they are challenging to exploit because it is difficult to precisely estimate the amount of blur and requires multiple images.

Finally, some state-of-the-art approaches explore what happens if patterns are deliberately introduced into the aperture, as illustrated in Fig. 13(b). As before, the captured image will still be

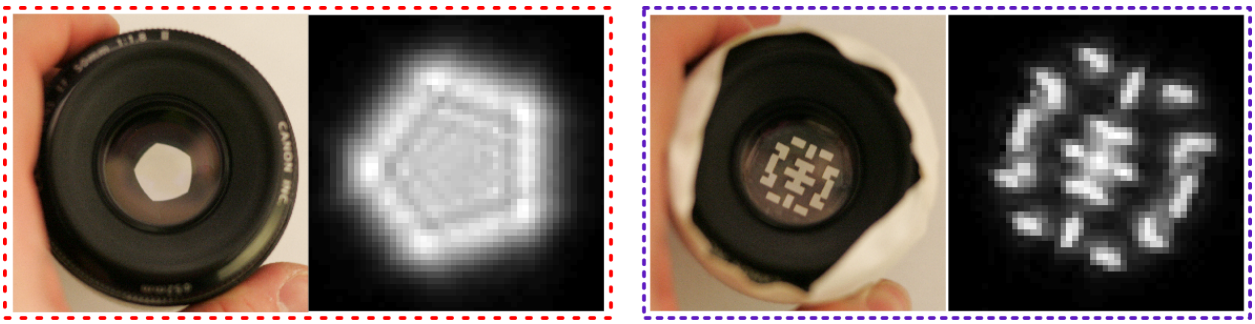

Figura 13. (a) Standard Canon 50 mm f /1.8 lens with the aperture partially closed and the resulting blur pattern. (B) Same camera but with a coded aperture inserted in the aperture and the resulting blur pattern(Levin et al., 2007).

blurred as a function of depth, with the blur being a scaled version of the aperture shape, but the aperture filter can be designed to discriminate between different depths.

2.5.2. Deep Optics. Since to the great advances in computational algorithms that have emerged in the last decade, deep learning can be used to design optical elements in conjunction with the weights of a convolutional neural network for a specific task (end-to-end optimization). Specifically, optimizing the parameters of optical elements and the point spread function have demonstrated high accuracy in areas such as extended depth-of-field, image super-resolution, and depth estimation in monocular vision systems, among others. Figure 14 illustrates the end-to-end approach proposed by Hayato et al. (Ikoma et al., 2021), in depth estimation and in-focus RGB reconstruction.

On the other hand, the optical element optimization approaches proposed above are mainly

---

[3] Image obtained from the project main page `https://www.computationalimaging.org/publications/deepopticsdfd/`.

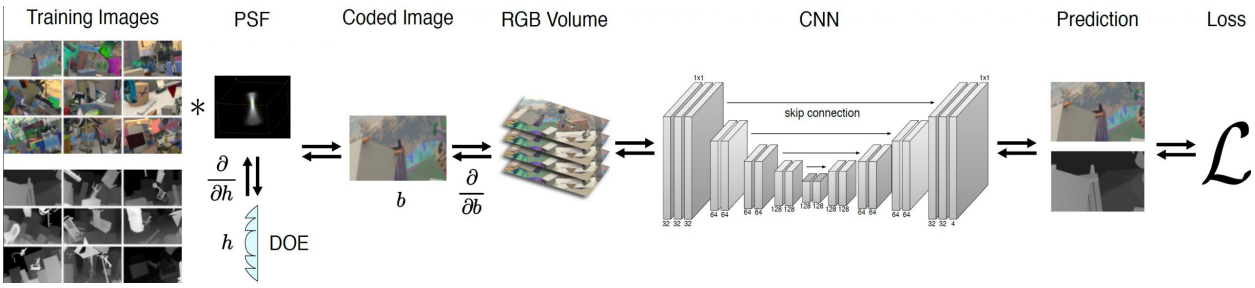

Figura 14. Illustration of an end-to-end (E2E) optimization framework. RGB and depth images of a training set are convolved with the depth-dependent 3D PSF created by a lens surface profile h. The resulting encoded sensor image is the input of the CNN. A loss function $\mathscr{L}$ is applied to the resulting RGB image and the depth map. The error is backpropagated into the CNN parameters and the surface profile.[3]

based on heuristic cost functions applied to PSFs, which may be a feasible approach for image deconvolution, but it remains unclear how a camera PSF affects higher-level computer vision tasks such as image classification; second, although image processing is applied to recorded images to remove residual aberrations or perform some inference tasks, the post-processing algorithm is usually independent of the optical design and does not provide meaningful information to guide it. Therefore, specific-purpose optical elements designed under an end-to-end optimization approach provide a significant advantage since the neural networks are trained with a wide variety of labeled data, which makes the designed optical element more versatile for use in specific environments.

For example, (Chang and Wetzstein, 2019) proposes an end-to-end design of the system optical elements and the recovery algorithm for the monocular depth estimation problem exploiting chromatic aberrations as depth-cues to encoded depth information while the image is captured. Specifically, the codification is made by a free-form lens jointly estimated with a CNN that recovers the depth map. Similar to (Chang and Wetzstein, 2019; Haim et al., 2018), Shiyu et al. exploited the emerging computational paradigm to design a phase mask to an extended depth of field in a

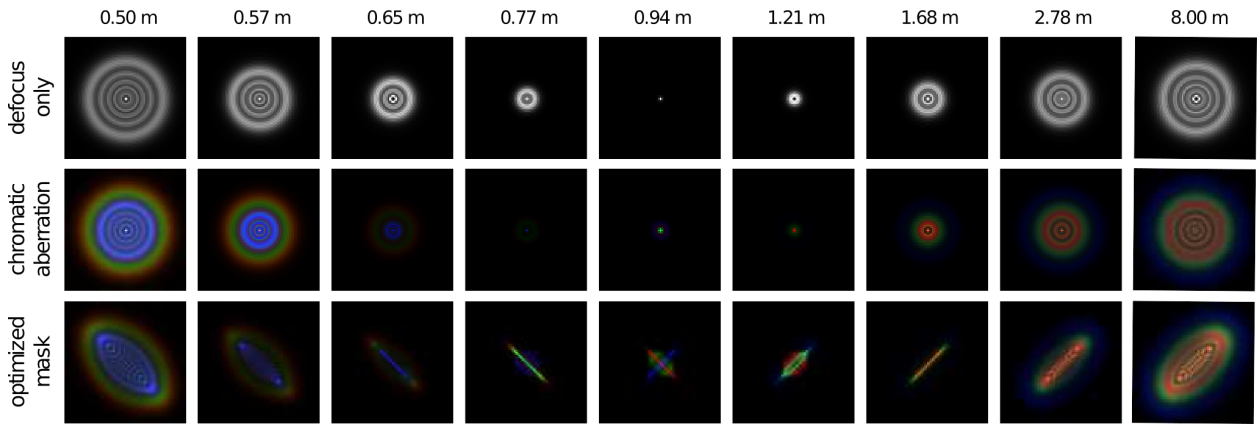

Figura 15. Point spread functions for different depths in an optical system. Traditional phase mask in the conventional thin lens (Top), a thin lens with defocus and chromatic aberrations (Middle), and optimized diffractive lens to promote defocus by chromatic aberrations by (Chang and Wetzstein, 2019) (Bottom).

conventional stereo system and finally obtain a more accurate disparity map. Common to all these approaches is to promote of optical depth cues. Meanwhile, these optical depth cues are obtained using an optical system with a depth-dependent point spread function. For example, Fig. 15 is depicted a set of different PSFs for different depths for a system, exploiting the relation between depth and chromatic aberrations (Chang and Wetzstein, 2019).

3. Proposed Method

This master thesis considers a stereo system that encodes depth information using a coded aperture. The performance of this system mainly depends on the coded apertures and the retrieval algorithm used to estimate the disparity map. Therefore, we propose jointly designing the stereo system and a neural network that functions as a decoder in an end-to-end (E2E) manner. As shown in Fig. 16, our proposed solution consists of two main components, 1) a differentiable optical layer, whose trainable parameters are the coded aperture values, 2) an encoder-decoder architecture, which is a combination of the U-Net (Ronneberger et al., 2015) network and the correlation module of the CNN proposed by Nikolaus et al. (Mayer et al., 2016) for disparity estimation, the union of these two models we call DispCorr-UNet. During training, the optical layer is fed with a pair of focused RGB images and their respective disparity map, thus generating the simulated encoded image. Next, the encoded image is provided to the DispCorr-UNet network, which produces the estimated disparity. Finally, the loss between the estimated disparity map and the real one is calculated to optimize the coded apertures and network weights using backpropagation theory. In the following subsections, we present the optical system model to obtain the simulated sensor measurement for each camera in a stereo configuration, using Fourier optics, the architecture of our CNN model, and implementation details.

3.1. Image Formation Model

To begin, we create a stereo model with two cameras and two single convex thin lenses with a focal length of f and a distance of z_i from the sensor, see Fig. 17. It is important to remark that,

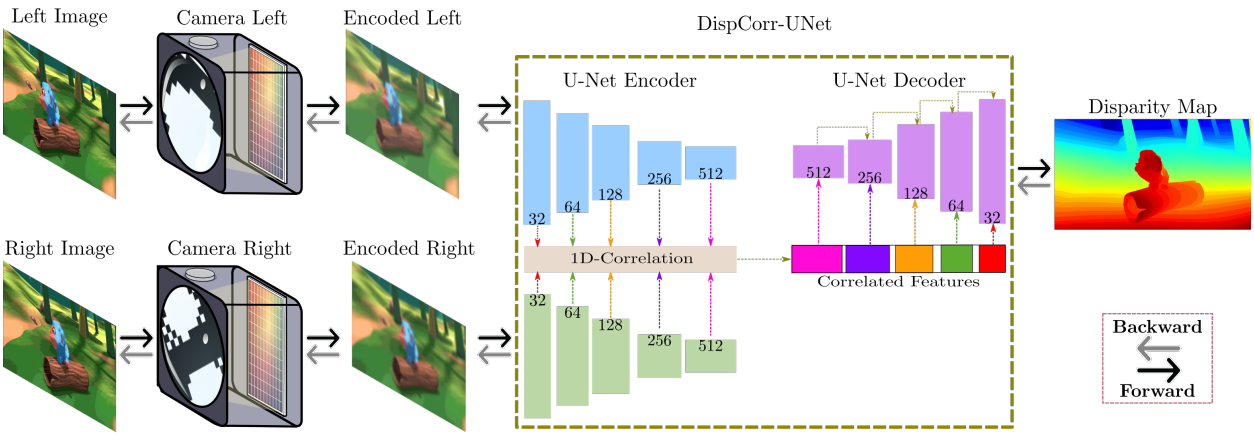

Figura 16. Our proposed architecture consists of two parts. In the optical layer, digital modeling of depth-dependent PSFs is performed from an optimized color aperture code, thus obtaining a depth-encoded image. In the reconstruction network, we use a U-Net based network for depth estimation from the coded image. The optical layer and reconstruction network parameters are optimized based on the defined loss between the estimated depth and the reference depth map.

as in conventional stereo systems, the two cameras must be the same; the mathematical modeling of the imaging system described throughout this section is the same for both cameras used in this work. Therefore, the thin-lens equation describes the relationship between in-focus distance and sensor distance:

$$\frac{1}{f} = \frac{1}{z} + \frac{1}{z_i}. \tag{9}$$

Most real-world scenes contain objects at various depths imaged with different PSFs. To simulate the PSF for every depth z, consider a point emitter of wavelength λ centered on the optical axis and located z away from the thin lens center. Then, the main idea is to propagate the light wave through the optical system to the sensor. We begin by sending a spherical wave from the point to the lens. As a result, an object at z in front of the lens is in focus at z_i behind the lens. The electric field of the spherical wave arriving at the lens is approximated using the Fresnel

approximation[4] as:

$$U_{\text{in}_{\lambda,z}}(x,y) = \exp\left[i\frac{k}{2}\frac{x^2+y^2}{z}\right]. \tag{10}$$

The wave front is then propagated through the lens by multiplying the input by the lens phase delay, $t(x,y)$:

$$t(x,y) = \exp\left[-i\frac{k}{2f}\left(x^2+y^2\right)\right]. \tag{11}$$

Since a lens has a finite aperture size, we insert an amplitude function $A(x,y)$ that blocks all light outside the open aperture. We also added the function $T(x,y,\lambda)$, which represents the coded aperture (CA) separated by z_t from the lens.[5] Assuming that the CA wavelength response remains approximately constant over a spatial squared region of size $\Delta_m \times \Delta_m$ named pixel, the CA can be modeled using a rectangular function as follows:

$$T(x,y,\lambda) = \sum_{i,j,k} W_{i,j,k}\text{rect}\left(\frac{x}{\Delta_m} - i, \frac{y}{\Delta_m} - j, \frac{\lambda}{\Delta_d} - k\right), \tag{12}$$

where $W_{i,j} \in [0,1]$ represents the discretized light response of the filter in the (i,j)-th position, and k-th band, and Δ_d represents the wavelength size of the discretized pixels (Bacca et al., 2021; Arguello et al., 2021). Then, multiply the amplitude, coded aperture, and phase modulation of the

---

[4] It assumes that the wavelength λ is significantly smaller than the travel distance z: $\lambda \ll z$.

[5] The proposed method takes into account the CA attached to the lens, i.e., $z_t \approx 0$.

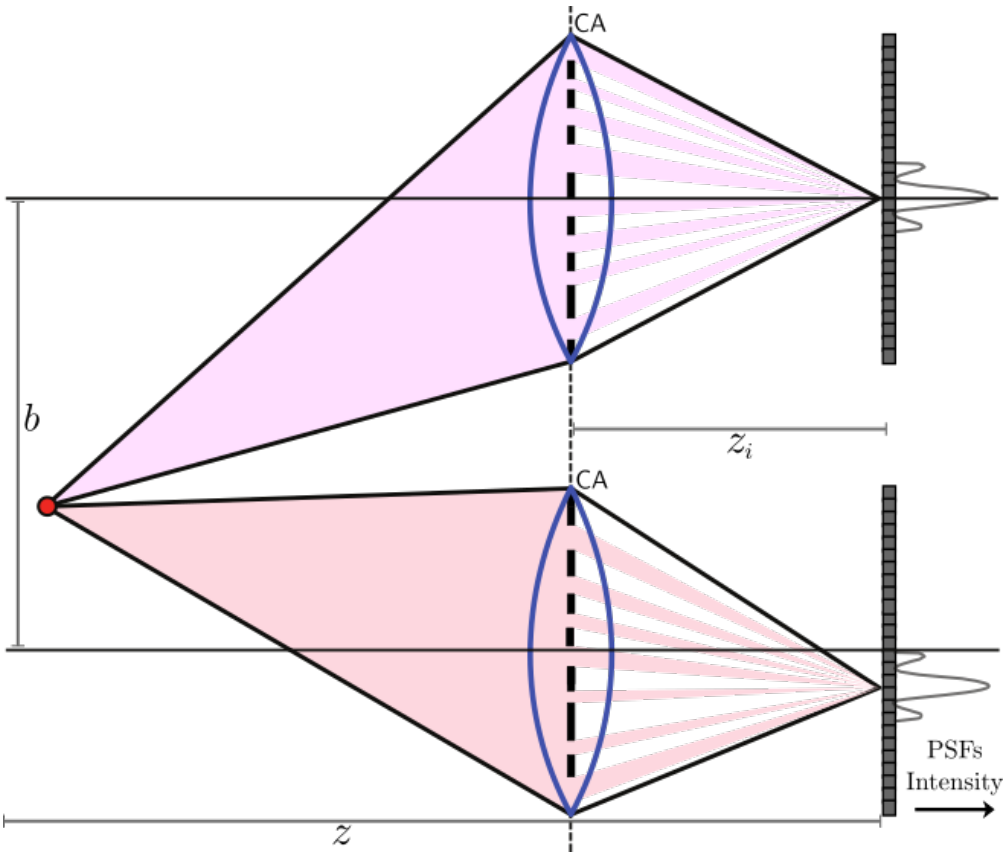

Figura 17. Optical propagation model of point sources through coded apertures in front of a thin lens.

lens with the input electric field to determine the electric field immediately after the lens and the coded aperture:

$$U_{\text{out}_{\lambda,z}}(x,y) = A(x,y)T(x,y,\lambda)t(x,y)U_{\text{in}}(x,y). \tag{13}$$

We propagate $U_{\text{out}_{\lambda,z}}(x,y)$ a distance z_i to the sensor using the exact transfer function (Goodman, 2005)

$$H_{\text{z}_\text{i}}(f_x,f_y) = \exp\left[ikz_i\sqrt{1-(\lambda f_x)^2-(\lambda f_y)^2}\right], \tag{14}$$

to finally obtain the field in the sensor

$$U_{\text{sen}\lambda,z}\left(x',y'\right) = \mathscr{F}^{-1}\left\{\mathscr{F}\left\{U_{\text{out}\,\lambda,z}(x,y)\right\}\cdot H_{z_{\text{i}}}\left(f_x,f_y\right)\right\}, \tag{15}$$

where (f_x, f_y) are spatial frequencies and $\mathscr{F}$ denotes the 2D Fourier transform. Since the sensor acquires light intensity, the point spread function (PSF) of the system is obtained using the magnitude-squared of (15):

$$\text{PSF}_{\lambda,z}\left(x',y'\right) = \left|U_{\text{sen}\lambda,z}\left(x',y'\right)\right|^2. \tag{16}$$

By following this sequence of forward calculations, we can generate a 2D PSF for each depth and wavelength of interest. For defocus blur cues, we calculate $U_{\text{out}\lambda,z}(x,y)$ for each color channel (Eq. 13), which results in different PSFs, that depends on the wavelength λ.

3.1.1. Depth-Dependent Image Formation Model. The simulated PSFs are employed to approximate the captured image by an RGB sensor of a given 3D scene using a layered representation that models the scene as a set of surfaces on discrete depth planes (Hasinoff and Kutulakos, 2007). This allows for the pre-computation of a fixed number of PSFs corresponding to each depth plane. We make a few modifications here to suit our data set consisting of pairs of all-in-focus RGB images and their discretized disparity maps. For an all-in-focus RGB image $\mathbf{I}$, a set of $j = 1 \ldots J$ discrete depth layers, and occlusion masks $\mathbf{M}_j$, the encoded image captured at wavelength λ is:

$$\hat{\mathbf{I}}_\lambda = \sum_{j=1}^{J} (\mathbf{I}_\lambda * PSF_{\lambda,j}) \odot \mathbf{M}_j, \tag{17}$$

where $*$ denotes 2D convolution for each color channel λ, and $\odot$ denotes element-wise multiplication. The cumulative occlusion masks $\mathbf{M}_j$ are alpha masks that modulate how much light from each layer is captured by the sensor (Chang and Wetzstein, 2019). The masks are generated with the blurred binary depth masks $\mathbf{A}_j$, from current and preceding layers:

$$\begin{aligned} \mathbf{M}'_j &= \left(1 - \mathbf{M}_{j+1}\right) \left(\mathbf{A}_j * \mathrm{PSF}_j\right), for \quad j < J \\ \mathbf{M}'_J &= \mathbf{A}_J * \mathrm{PSF}_J \,. \end{aligned} \tag{18}$$

Here, layer J is the layer closest to the camera that is not occluded by any additional layers, so $\mathbf{M}'_J$ consists of the regions of the depth map that fall into layer $\mathbf{A}_J$, blurred by PSF_J (Chang and Wetzstein, 2019). The layers occlude each layer behind J in front of it. Finally, $\left\{\mathbf{M}'_j\right\}$ are normalized to $\left\{\mathbf{M}_j\right\}$ such that the sum of occlusion mask weights at each pixel location sums to 1. Finally, the RGB channels of the captured encoded image can be modeled as:

$$\ddot{\mathbf{I}}_c = \hat{\mathbf{I}}_\lambda \cdot \mathscr{R}_c(\lambda), \tag{19}$$

where $\mathscr{R}_c(\lambda)$ represents the sensor response in the c-th channel.

Figura 18. Visual representation of the 1D-correlation module, between a stereo pair of images.

3.2. Disparity Reconstruction Network

We use an encoder-decoder scheme based on convolutional neural networks. The encoder is in charge of simulating the optical process of acquiring the stereo pair images (optical layer). The decoder is in charge of extracting the encoded depth information in the output images. Specifically, the DispCorr-UNet is composed of two different U-Net encoder that extracts the features of the left and right images separately, the features extracted by each U-Net-encoder module are correlated by the *1D-Correlation* module proposed in (Mayer et al., 2016), then the correlation $\mathscr{F} \in \mathbb{R}^{c\times w\times h}$ between two left and right features maps $f_l, f_r \in \mathbb{R}^{c\times w\times h}$, can be described as:

$$\mathscr{F} = \frac{1}{D}\sum_{d=0}^{D-1} f_l(x,y) \odot f_r(x-d,y), \tag{20}$$

where D is the number of possible disparity values (Mayer et al., 2016), and $\odot$ represent the scalar product, a visual representation of this 1D-correlation scheme is shown in Fig. 18. Once we have the correlated features, they are passed to a traditional U-Net decoder since it is widely used for pixel-level prediction and is indispensable for disparity estimation. This scheme is illustrated in Fig. 16.

3.3. Implementation Details

Once we have defined our image formation model and the architecture of our proposed network, we implement computational simulations. The model was implemented and trained in Python 3.8, using the automatic differentiation artificial intelligence module Pytorch. The training was performed on a Linux computer equipped with an Intel Xeon W-3223 processor with a frequency of 3.50 GHz and an NVIDIA GeForce RTX 3090 graphics card with 24 GB memory.

3.3.1. Dataset. Our end-to-end approach is trained on a synthetic dataset called Scene-Flow, which consists of dense ground truth disparity maps with $35,454$ images for training and $4,370$ for testing (Mayer et al., 2016). Since this dataset is synthetic and its generation process is not entirely accurate, we reviewed it. We found scenes containing outliers in their disparity map, Fig. 19, which negatively influences our end-to-end optimization approach. Therefore, we filtered this dataset and discarded scenes containing a disparity map outside the range of $[0-192]$; this range was chosen according to the state-of-the-art, leaving a total of $27,225$ and $4,008$ image pairs for training and testing respectively. Finally, to increase the variability of the data during training, a random harvest of size 256 times 256 was used.

3.3.2. Loss Function. To optimize all the parameters of our model, the following cost functions were taken into account:

$$\mathscr{L} = \alpha\mathscr{L}_s + \gamma\mathscr{L}_{cl} + \gamma\mathscr{L}_{cr}, \tag{21}$$

where $\mathscr{L}$ is the total loss and $\alpha, \gamma \in \mathbb{R}$ are weighting coefficients for each loss. The descrip-

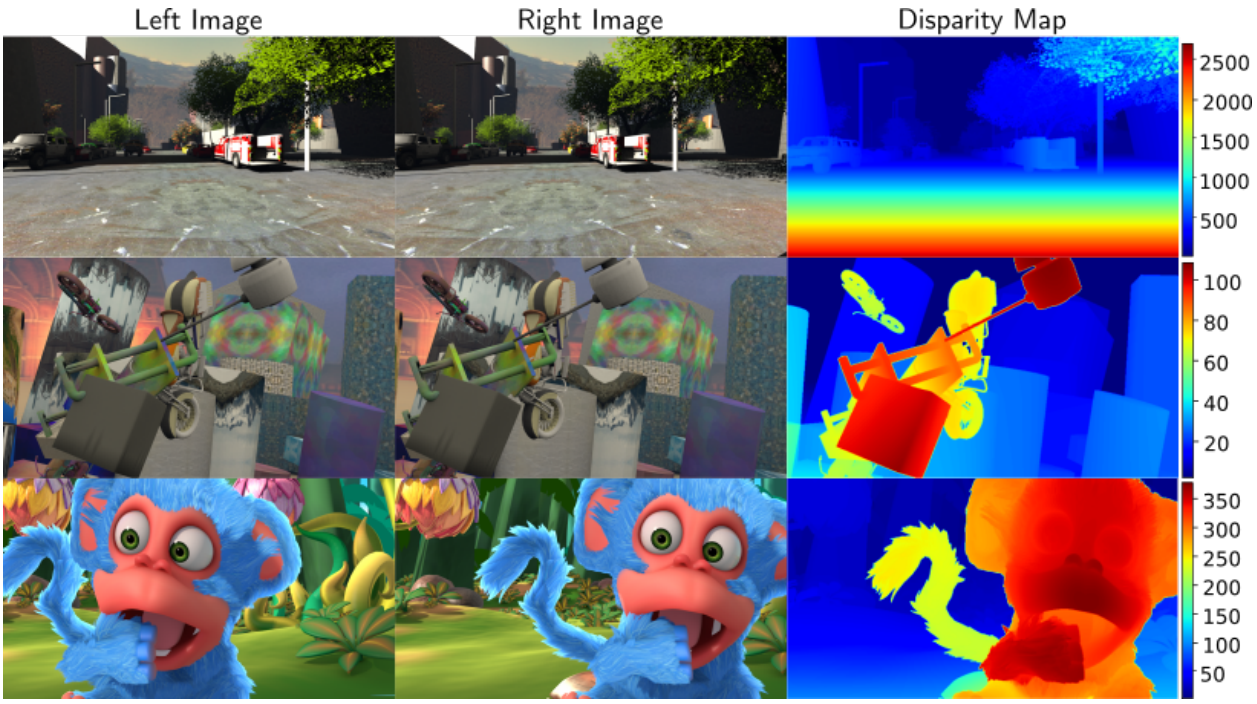

Figura 19. Some stereo pairs images of Scene-Flow dataset.

tion of each term of the total loss function is explained below:

- $\mathscr{L}_s$: Because the network can make disparity estimates, which can be far from their ground-truth value, we chose the cost function called L_1 Smooth (Girshick, 2015), which is a combination of the ℓ_1 and ℓ_2 norm, this cost function is represented as follows:

$$\mathscr{L}_s(x) = \frac{1}{K}\sum \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{22}$$

where x is the difference between the ground truth and the estimated disparity map by the DispCorr-UNet.

- $\mathscr{L}_{cl}$ and $\mathscr{L}_{cr}$: To optimize the left and right coded apertures over an end-to-end approach, it is necessary to include a regularization term in the cost function (Bacca et al., 2021), which promotes particular properties in the disparity estimation task:

$$\mathscr{L}_{c^*} = \frac{1}{K}\sum \Phi_*^2(\Phi_* - 1)^2, \tag{23}$$

where, Φ_* represent the left or right coded aperture and K is the total pixels of the apertures.

3.3.3. Evaluation Metrics. Following previous state-of-the-art works, we adopt the following evaluation metrics to assess the performance of our disparity prediction model quantitatively. Specifically, we use:

- Root Mean Squared (RMSE): $\sqrt{\frac{1}{N}\sum_i^N |y_i - \tilde{y}_i|^2}$
- End Point Error (EPE): $\frac{1}{N}\sum_i^N \|y_i - \tilde{y}_i\|$
- D1: $\frac{1}{N}\sum_{i=1}^N min\left(\|y_i - y_i^*\| > 3, \|y_i - y_i^*\| > y_i * 0.05\right)$
- Thresholded Accuracy (δ_j): $\frac{1}{N}\sum_i^N \text{máx}\left(\frac{\tilde{y}_i}{y_i}, \frac{y_i}{\tilde{y}_i}\right) < 1.25^j$

where y_i and $\tilde{y}_i$ are ground-truth and estimated disparity map, respectively, N is the total pixels and $j = \{1,2,3\}$. Smaller values on RMSE, EPE, and D1 error are better and higher values on δ_j threshold are better.

4. Simulations and Results

Once we have defined our model, the cost function, and the data set to be used, we perform computational simulations to validate our approach accuracy. We use the Adam optimizer during the training with its default parameters, and empirically, we find that using different learning rates for the optical encoder and DispCorr-UNet improves performance. Therefore, in all the simulations shown in this section, a learning rate of $5e^{-2}$ and $5e^{-4}$ was used for the optical encoder and the network, respectively, with a batch size of 32 for 100 epochs. The regularization parameters of the cost function were assigned a value of 1.5 for α, and for γ they were assigned a variable increment over the epochs, from $1e^{-9}$ to $1e^{-1}$ as proposed Jorge et al. (Bacca et al., 2021).

4.1. Ablations Studies

To demonstrate the accuracy of our approach at the optical encoding level using coded apertures and the effectiveness of our CNN in disparity estimation, we have developed the following experiments:

- **Experiment 1:** The optical layer is removed, and the original images from the Scene-Flow dataset are used to estimate the disparity with DispCorr-UNet.
- **Experiment 2:** A conventional $F8$ lens with the depth-dependent image formation model mentioned in the previous section encodes the depth information.
- **Experiment 3:** Our proposed approach, by including the optical layer frozen and using a random binary coded aperture.

- **Experiment 4:** Our proposed approach, by including the optical layer frozen and using a random color coded aperture.
- **Experiment 5:** The proposed end-to-end optimization approach to design a binary code aperture.
- **Experiment 6:** The proposed end-to-end optimization approach to design a color code aperture.
- **Experiment 7:** Sharing color coded aperture optimized under an end-to-end optimization approach for the left and right camera.
- **Experiment 8:** End-to-end optimization over a color coded aperture with additional refinement module Spatial Propagation Network (Liu et al., 2017).
- **Experiment 9:** Replace disparity correlation module in DispCorr-UNet by a single addition module.
- **Experiment 10:** Possible disparity parameter D in the Eq. 20 with a value of 3 throughout the correlation module.
- **Experiment 11:** Possible disparity parameter D in the Eq. 20 with a value of 5 throughout the correlation module.
- **Experiment 12:** Possible disparity parameter D in the Eq. 20 with a value of 7 throughout the correlation module.

Experiment	**RMSE**↓	**EPE**↓	**D1**(%) ↓	**3px**(%) ↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
1	4.9737	2.1451	11.783	15.765	0.9641	0.9851	0.9918
2	4.8916	2.3044	12.641	15.525	0.9520	0.9805	0.9890
3	3.7560	1.4265	7.7512	10.915	0.9748	0.9899	0.9941
4	3.8652	1.4504	8.0593	10.865	0.9731	0.9894	0.9940
5	2.6969	1.1052	5.2108	8.1927	0.9871	0.9955	0.9973
6	**2.2821**	**0.9453**	**4.1681**	**6.4860**	**0.9899**	**0.9967**	**0.9980**
7	4.5662	2.0148	13.753	14.618	0.9593	0.9833	0.9895
8	4.2954	1.6921	9.3734	11.764	0.9553	0.9792	0.9873
9	3.3629	1.3348	6.8634	12.451	0.9810	0.9918	0.9948
10	3.0126	1.3530	7.8317	11.778	0.9830	0.9936	0.9949
11	3.2346	1.2958	6.7284	9.8668	0.9817	0.9926	0.9952
12	3.5688	1.4069	8.8246	10.536	0.9759	0.9902	0.9930
13	3.7492	1.3615	8.9678	11.546	0.9781	0.9901	0.9935

Tabla 1
Ablation studies results, the best results are in bold and the second best are underlined.

- **Experiment 13:** Incremented linearly for the possible disparity parameter D in Eq. 20, from 1 to 5 in the correlation module.

As can be seen in the above table, Tab 1, 13 experiments were performed, in which different configurations, both optical and software, were carried out to find the best architecture to estimate the disparity in a stereo camera system. Initially, we want to mention that state-of-the-art stereo systems traditionally use the metrics 3px, D1, and EPE to quantitatively measure a model's performance in disparity estimation. However, we decided to use other metrics used in monocular-depth vision systems, such as RMSE and Threshold accuracy, to compare better the experiments performed.

The first two experiments show that using traditional cameras is not significantly contributing to disparity estimation since these two experiments report the worst metrics compared to the others, in which some type of optical encoding is included using coded apertures. Traditionally in state-of-the-art, when coded apertures are used in monocular or stereo vision systems for depth estimation, they are usually binary. Nevertheless, it has been shown that in areas such as spectral imaging (Galvis et al., 2019), monocular-depth estimation (Paramonov et al., 2016), compressive sensing (Arguello and Arce, 2014), among others, using color coded apertures (CCA) increases the accuracy of the desired task. Therefore, experiments 3 to 6 were performed to evaluate the difference in performance using binary and color code apertures. In experiment 4, it is evident that using a random CCA does not improve the accuracy compared to experiments 3 and 5, which use binary coded aperture. However, in experiment 6, it is evident that optimizing a CCA under an end-to-end optimization approach increases the accuracy in the estimation of disparity compared to all previous experiments performed; for this reason, in subsequent ablation studies, a CCA is always used.

Since the proposed network uses two feature extractors, and these are the key to the disparity estimation, the hypothesis arose in which it was expected that if the coding were the same, the feature extractors would be more accurate and thus obtain better accuracy. However, as observed in experiment 7, this approach did not improve the accuracy compared to the experiments conducted so far; on the other hand, most of the unsupervised approaches mentioned in subsection 2.4 and some deep learning approaches once they estimate the disparity map, proceed to use a refinement algorithm to improve the texture of this map, so to test this approach, experiment 8 was

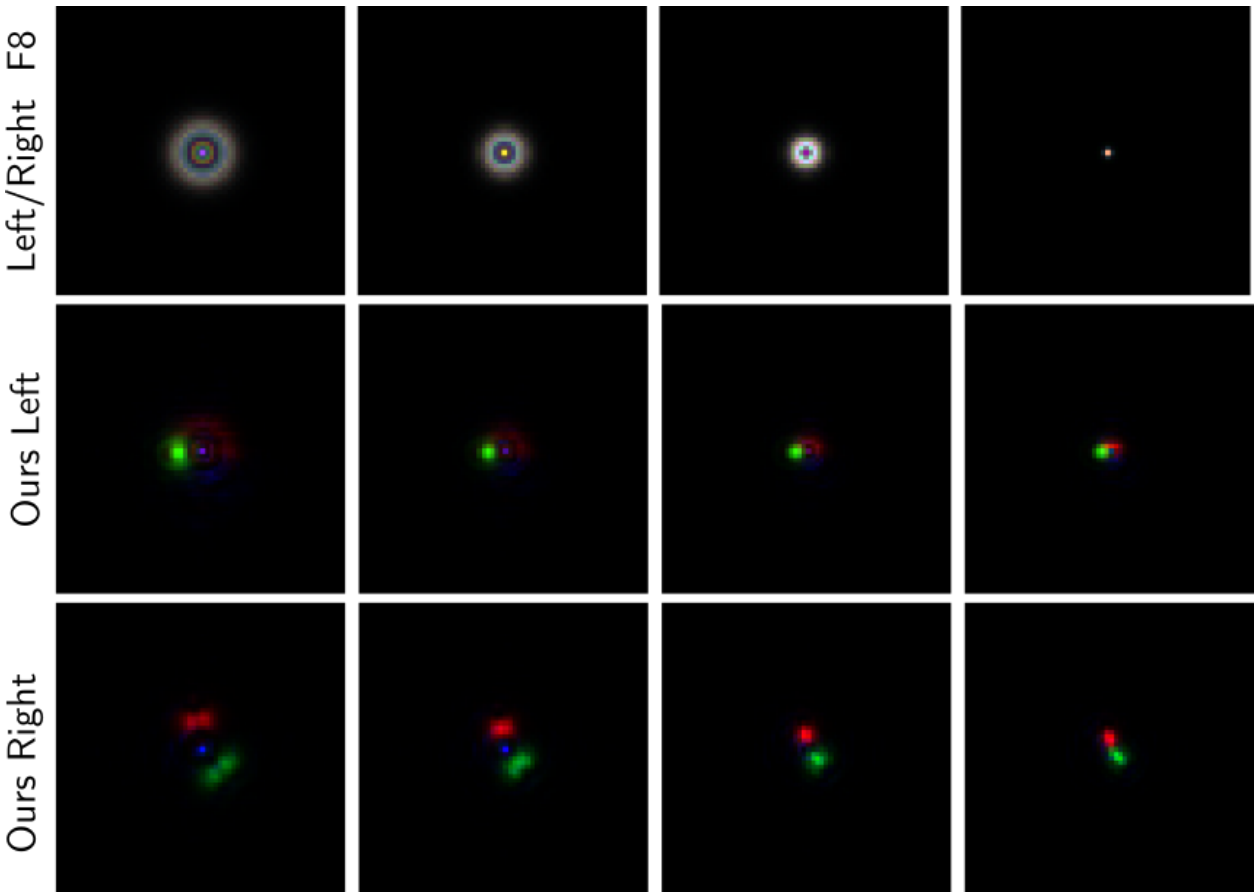

Figura 20. Simulated PSFs for conventional F8 lens and optimized system in experiment 4.

performed, where it is evident that using an additional refinement module, does not improve the accuracy, on the contrary, worsens it, this is believed that because our architecture is based on UNet and this architecture has high accuracy at the pixel level, and therefore the refinement module is unnecessary.

To show that the correlation module plays an important role, in experiment 9, we eliminate this module; for a simple approach, how is the addition of the characteristics maps, this is because the correlation shown in Fig. 18 is visually seen how in the overlap between the left and right image, and a similar one of being able to see this overlap, it would be through addition, but how

is show quantitatively, use the correlation module obtains greater precision. On the other hand, in experiments from 1 to 8, the possible disparity D, used in Eq. 20, has had a value of 1; therefore, in the remaining experiments (10-13), it is analyzed how this variable influences the precision of the model, and it is evident that with the value of 1, the most incredible precision is obtained, which is obtained in the other experiments. Finally, in Fig. 20, the PSFs from experiments 2 and 4 are shown, where it is evident that our optimized PSFs for the left and right camera have variations at different depths, providing complementary depth cues to aid disparity prediction in problem areas compared to the Airy disk of the conventional F8 lens.

4.2. Analysis and Evaluation

Once the ablation studies have been performed, demonstrating our workflow, we compare our approach with state-of-the-art methods. Specifically, as mentioned in the subsection 2.4.4, deep learning models are not optimized under an end-to-end approach (non-E2E), and those that are (E2E). Therefore, we proceed to compare those two approaches. In the upper part of Tab. 2, we show results from different state-of-the-art non-E2E works, which receive a pair of stereo images in focus for disparity estimation. As seen in this table, our E2E method does not obtain superior results compared to other works since they somehow seek, through computationally expensive software techniques, to extract as much information as possible, which is not the strength of our work. For example, the network that achieves the best performance is the network proposed by Google, which is based on extracting compact feature representations, on a disparity-based initialization of the network weights, and finally on a slanted window refinement model. As seen in these networks, most achieve high accuracy in disparity estimation from software, not hardware solutions. Howe-

Network	EPE	D1(%)	3px(%)
MobileStereoNet (Shamsafar et al., 2022)	0.80	6.15	7.06
HITNet L (Tankovich et al., 2021)	0.43	4.70	2.57
HITNet XL (Tankovich et al., 2021)	**0.36**	4.09	**2.21**
AANet (Xu and Zhang, 2020)	0.87	9.30	-
AANet + (Xu and Zhang, 2020)	0.72	7.4	-
LEAStereo (Cheng et al., 2020)	0.78	7.82	-
WStereo (Garg et al., 2020)	0.70	7.70	2.98
GA-Net (Zhang et al., 2019)	0.84	9.90	-
SegStereo (Yang et al., 2018)	1.45	**3.5**	-
DispNetCorr1D (Mayer et al., 2016)	2.33	10.0	-
AnyNet (Wang et al., 2019)	4.27	22.6	24,8
GC-Net (Kendall et al., 2017)	1.84	9.67	-
CRL-Net (Pang et al., 2017)	1.67	6.7	-
LD-Net (Liang et al., 2018)	1.27	4.9	-
DSM-Net (Zhang et al., 2021)	0.761	8.31	4.07
DispSharpNet (Tan et al., 2021)	1.52	-	7.85
AnyNet-E2E	3.94	19.6	21,3
Our Binary	1.11	5.21	8.19
Ours Color	0.95	4.18	6.65

Tabla 2
State-of-the-art comparison, the first methods are traditional approaches using RGB in-focus stereo images as input, and the last approaches are with optical encoding.

ver, it still outperformed some state-of-the-art works, such as methods DispNetCorr1D, GC-Net, LD-Net, etc.

On the other hand, to compare ourselves against E2E methods, we searched the state-of-the-art for works in which some kind of optical design for depth estimation in stereo systems has been performed. Only the work proposed by Shiyu et al. (Tan et al., 2021) has been found, in which they optimize phase masks to increase the depth of field and thus obtain a depth encoding

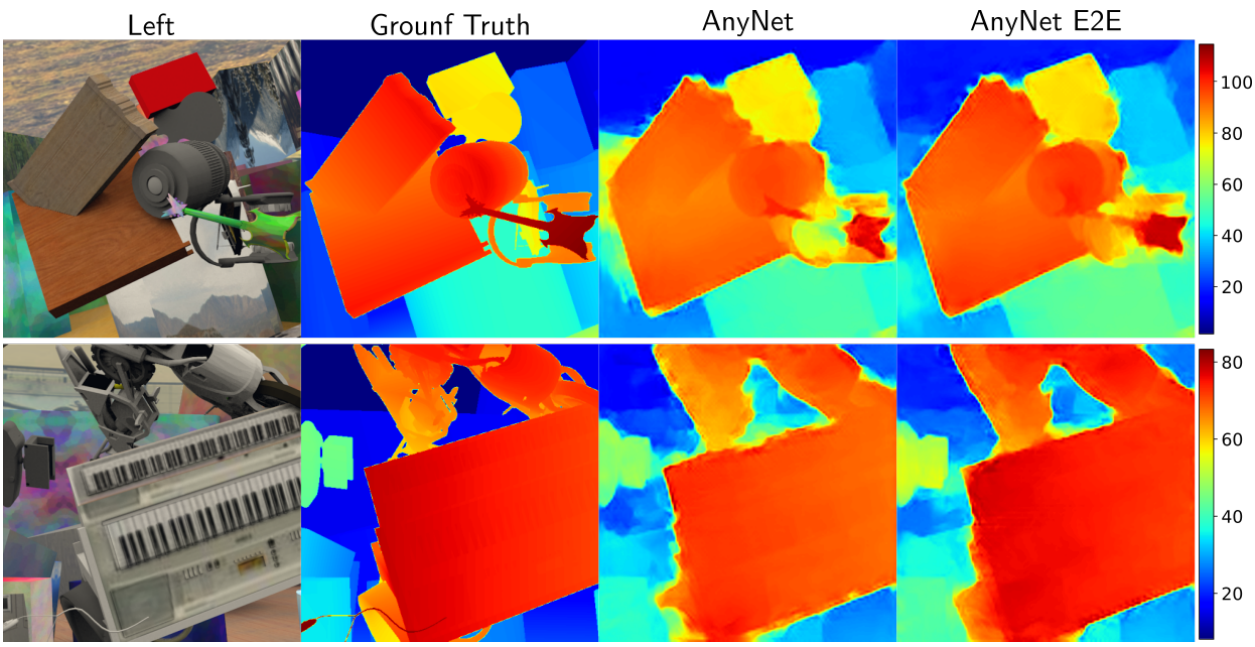

Figura 21. Disparities predictions of original AnyNet and our E2E AnyNet.

in a stereo system. Due to the scarcity of E2E approaches and to show a higher validity of our approach, we decided to couple our optical layer to another state-of-the-art network, different from the model proposed by us, to show that our optical encoding approach can be adapted to any state-of-the-art network for disparity estimation and obtain better accuracy than the original network. Specifically, to validate this hypothesis, we have used the AnyNet network, designed to quickly and accurately perform depth estimation. Because the original paper uses the Scene flow dataset for initial training of the network but does not report evaluation metrics in the test set, AnyNet was trained by us from scratch for 100 epochs, with a learning rate of $5e^{-3}$ and a mini-batch of 32, and subsequently trained from scratch again under the E2E optimization approach with the same learning parameters as the previous experiment and a learning rate of $5e^{-4}$ for the optical layer. The results of this experiment are shown in Tab. 2.

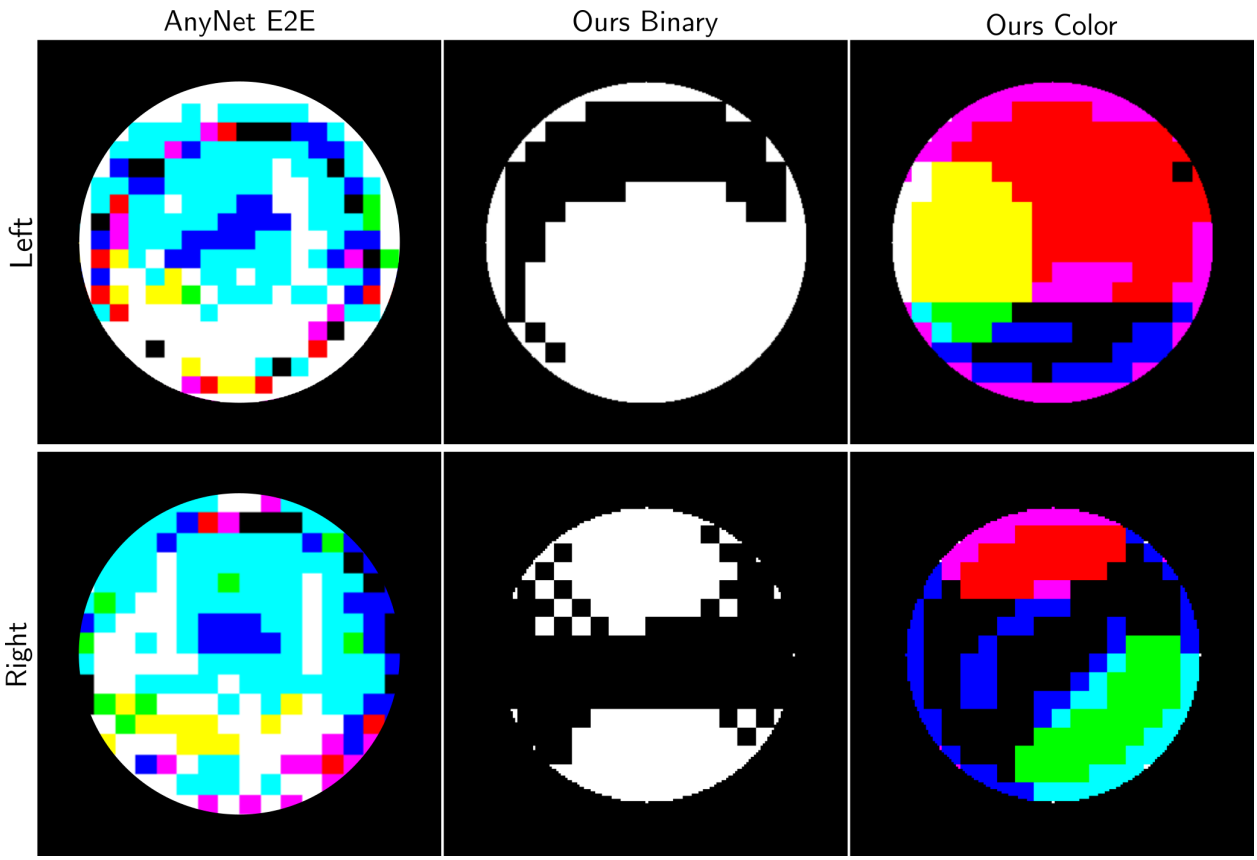

Figura 22. Optimized coded apertures over our E2E approach.

As reported in Tab. 2, the results we obtained on AnyNet are not the most favorable for this model because we did not adequately explore hyperparameters (learning rate, mini-batch size). We decided to leave the same hyperparameters used for our model, but even these results manage to demonstrate the hypothesis stated above. As reported by the visual and quantitative results shown in Fig. 21 and Tab. 2, respectively, and the results shown from experiments 1 and 2 of the ablation studies, see Tab. 1, employing optical encoding by using coded apertures improve the accuracy of depth estimation in a stereo system.

Fig. 22 shows the color aperture optimized under an E2E optimization approach using Any-

Net, and the binary and color apertures optimized by our DispCorr-UNet network. The left binary aperture optimized by us has a similar but rotated structure to one of the apertures designed by Zhou et al. (Zhou et al., 2011) using genetic algorithms for depth estimation using defocus techniques. For us, the similarity between these apertures is a further indicator that our approach is successful since this work explores the design of binary CA and their impact on depth estimation in a monocular vision system. Finally, in Fig. 23, we show some disparity maps obtained by the three E2E optimization approaches addressed in this work, where the high accuracy of our approach in conjunction with the proposed network, either using a binary or color aperture, is evident.

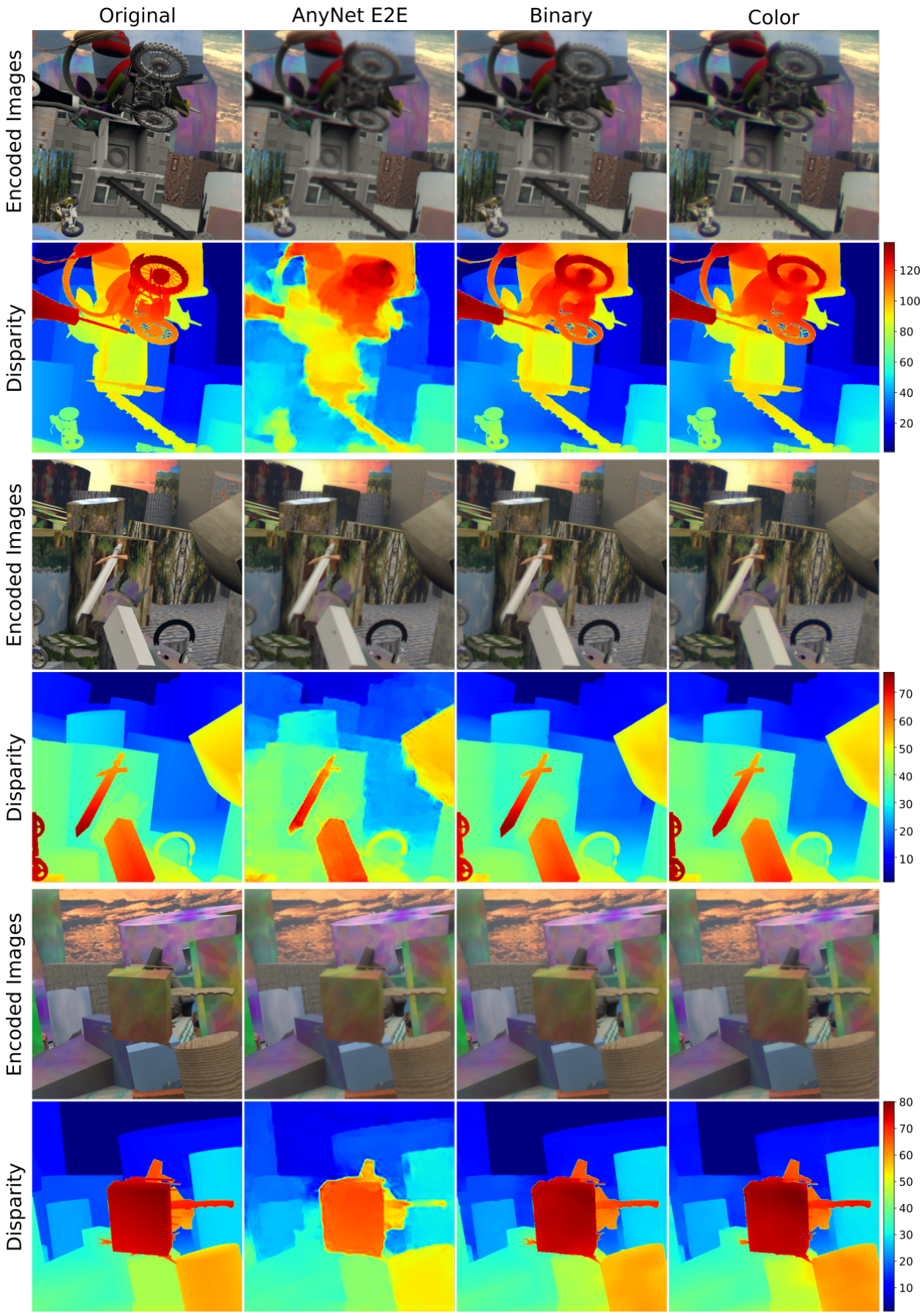

Figura 23. Visual results of end-to-end models, in test set Scene Flow dataset.

5. Real Hardware Implementation

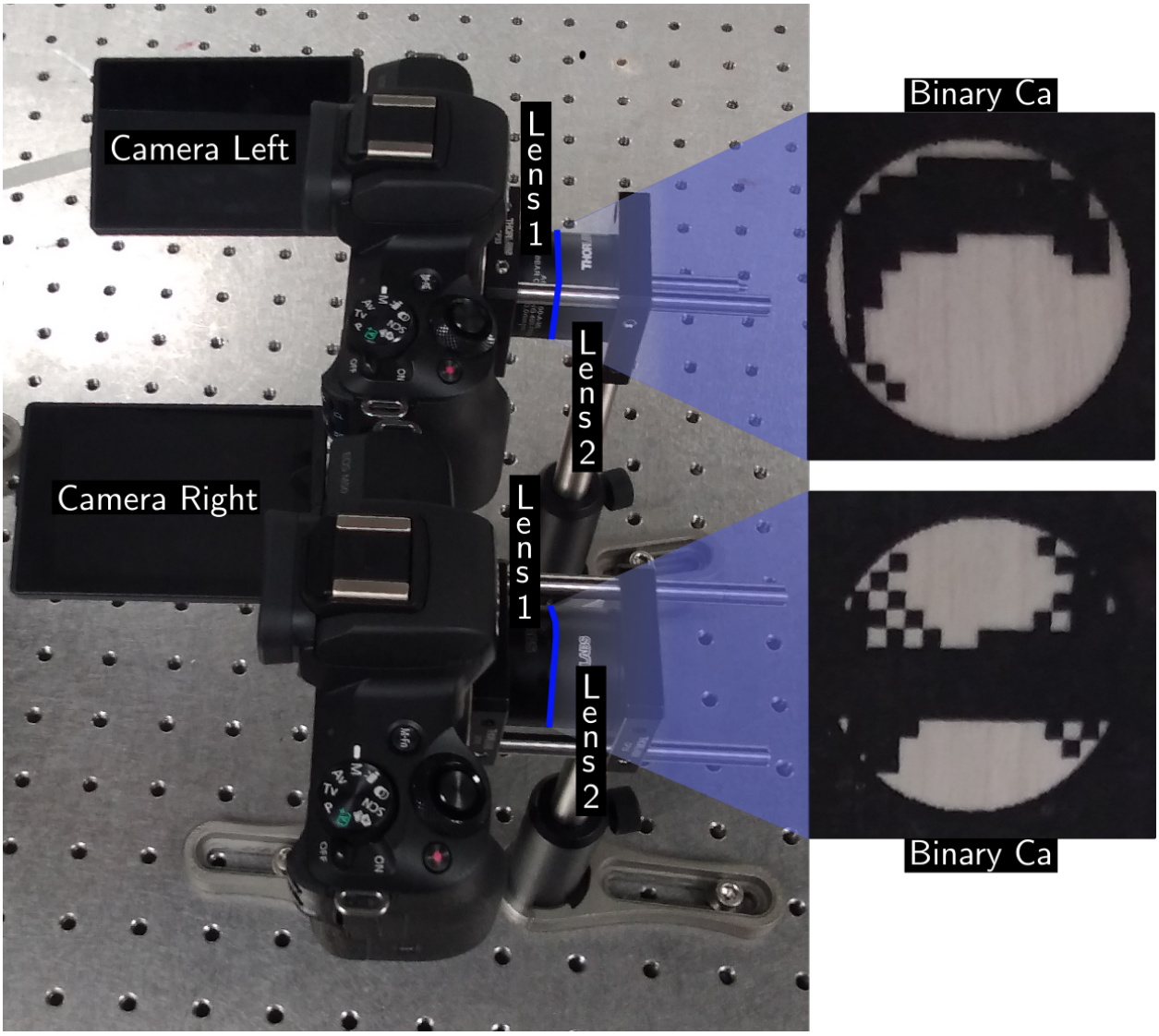

Figura 24. Real hardware implementation of the proposed approach for disparity estimation with our optimized binary coded apertures.

To demonstrate our proposed approach, we try to build a real implementation similar to the system shown in Fig. 17; for this, we located an optimized coded aperture in the middle of a

composed lens, where lens 1 and 2 have a focal length of 100 and 50 millimeters, respectively. This composed lens is attached to a Canon camera (EOS M50) with a 24.1 megapixel CMOS sensor; see Fig. 24. To fabricate our binary code aperture, it has been printed on transparent films using an HP-LaserJet-M608 printer. Once the system is implemented, the performance of the proposed method may be affected by implementation mismatch and CA fabrication errors. Therefore, we capture a series of images from a point white light source to calibrate the simulated system with the captured PSFs, see Fig 25. Next, an adjustment to the disparity estimation network is necessary. More specifically, we obtain the PSFs from the real system and retrain the estimation network for 100 epochs with a learning rate of $3e-5$.

Figura 25. Simulated and laboratory PSFs for the real hardware implementation.

Once our disparity estimation network has been fine-tuned with the PSFs obtained in the laboratory, we proceed to capture a set of stereo images to evaluate the accuracy of our method. As can be seen in Fig. 26, our approach has been evaluated on three pairs of stereo images captured by the system shown in Fig. 24; in the first two rows, we show scenarios in which our approach obtains accuracy according to the depths observed in the captured images, on the other hand, in the third row we can observe that the disparity estimation is not the most adequate, In essence,

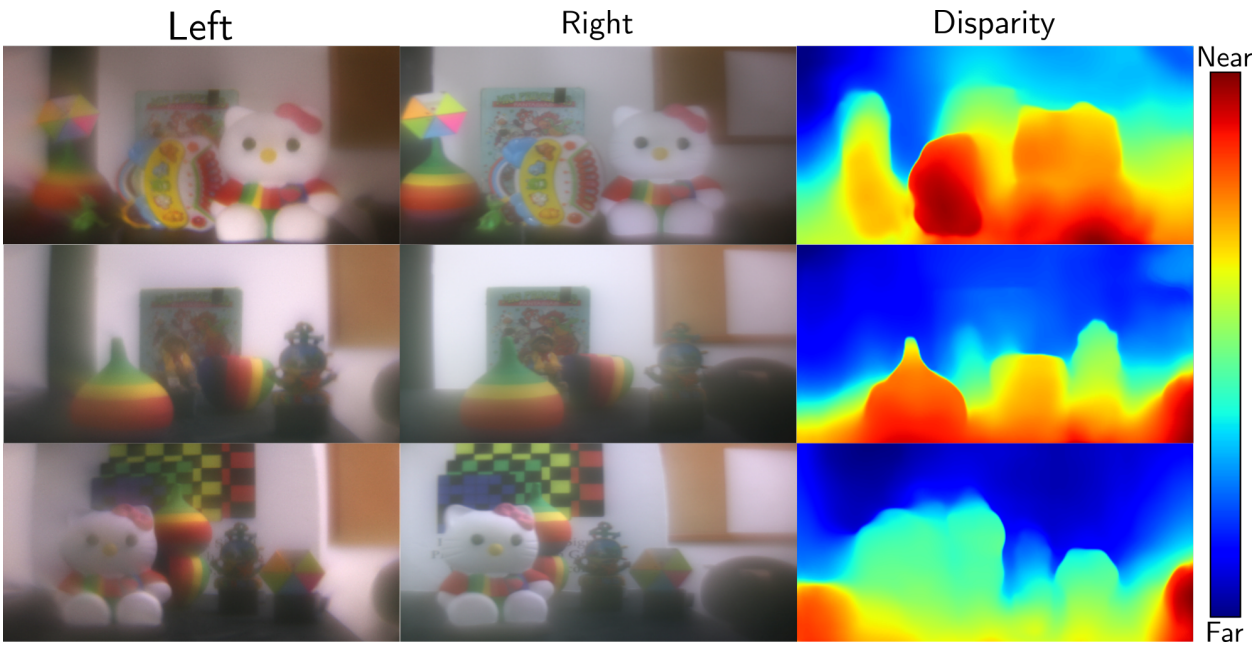

Figura 26. Experimental disparity estimation of various scenes using our real hardware prototype.

the network assumes that almost all the objects present in the scene belong to the same depth, we suspect that this case is due to two reasons, the first is due to errors in the implementation of the system, and the second is due to the high occlusion present in the scene, as it is well known in state of the art, stereo systems have difficulties in estimating the depth in areas of high occlusion. Finally, the disparity maps obtained from the natural scenes correspond to a relative depth, not absolute, because the data set used for training and fine-tuning the network is synthetic and does not correspond to real measures.

6. Conclusions

This research proposes an optical coding system for low-cost stereo systems that can retrieve high-resolution 3D information. The proposed system consists of introducing a coded aperture into the aperture plane. The coded aperture pattern is learned jointly with a disparity estimation network under an end-to-end optimization approach. The optimized coded apertures create a point spread function dependent on the depth at which the objects in the scene are located. This allows us to retrieve a more accurate disparity map than a conventional camera system. Additionally, we have proposed a new convolutional neural network derived from UNet in conjunction with a feature correlation module for disparity estimation.

We demonstrate through different ablation experiments that the proposed stereo system employing optimized coded aperture codes improves the depth estimation task by up to 4.1 % and 9.2 % in D1 and 3px errors metrics, respectively, compared with a stereo system without additional optical codification. We also show that our approach outperforms state-of-the-art works, which rely only on software techniques or hardware. Additionally, we show that our coding approach can be adapted to any state-of-the-art models and increase the accuracy up to 3.5 % in the 3px error metric, compared to its original approach (e.g. AnyNet CNN), obtaining, as a final product, a low-cost optical coding system compared to other methods that employ phase masks or diffractive optical elements. Additionally, using an experimental prototype, we have demonstrated the effectiveness of our system in depth recovery in natural scenes.

Referencias Bibliográficas

Arguello, H. and Arce, G. R. (2014). Colored coded aperture design by concentration of measure in compressive spectral imaging. *IEEE Transactions on Image Processing*, 23(4):1896–1908.

Arguello, H., Pinilla, S., Peng, Y., Ikoma, H., Bacca, J., and Wetzstein, G. (2021). Shift-variant color-coded diffractive spectral imaging system. *Optica*, 8(11):1424–1434.

Bacca, J., Gelvez-Barrera, T., and Arguello, H. (2021). Deep coded aperture design: An end-to-end approach for computational imaging tasks. *IEEE Transactions on Computational Imaging*, 7:1148–1160.

Barnard, S. T. (1987). Stereo matching by hierarchical, microcanonical annealing. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER.

Barnard, S. T. and Fischler, M. A. (1982). Computational stereo. *ACM Computing Surveys (CSUR)*, 14(4):553–572.

Bhoi, A. (2019). Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*.

Boric, S., Schiebel, E., Schlogl, C., Hildebrandt, M., Hofer, C., Macht, D. M., et al. (2021). Research in autonomous driving–a historic bibliometric view of the research development in autonomous driving. *International Journal of Innovation and Economic Development*, 7(5):27–44.

Bradski, G. and Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. .º'Reilly Media, Inc.".

Calderón Saavedra, D. S. (2012). Generación de mapas de profundidad a partir de imágenes estéreo utilizando registro no rígido.

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.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al. (2019). Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757.

Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., and Ge, Z. (2020). Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169.

Dornaika, F. (1995). *Contributions à l'intégration vision robotique: calibrage, localisation et asservissement*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*.

Forsyth, D. A. and Ponce, J. (2002). *Computer vision: a modern approach*. prentice hall professional technical reference.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011.

Fusiello, A., Roberto, V., and Trucco, E. (1997). Efficient stereo with multiple windowing. In *Proceedings of IEEE Computer Society conference on computer vision and pattern recognition*, pages 858–863. IEEE.

Galvis, L., Mojica, E., Arguello, H., and Arce, G. R. (2019). Shifting colored coded aperture design for spectral imaging. *Applied optics*, 58(7):B28–B38.

Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q., and Chao, W.-L. (2020). Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, 33:22517–22529.

Gehrig, S. K., Eberli, F., and Meyer, T. (2009). A real-time low-power stereo vision engine using semi-global matching. In *International Conference on Computer Vision Systems*, pages 134–143. Springer.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Goodman, J. W. (2005). *Introduction to Fourier optics*. Roberts and Company Publishers.

Grisetti, G., Kümmerle, R., Stachniss, C., and Burgard, W. (2010). A tutorial on graph-based slam. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43.

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.

Hannah, M. J. (1988). Digital stereo image matching techniques. *International Archives of Photogrammetry and Remote Sensing*, 27(B3):280–293.

Haralick, R. M. and Shapiro, L. G. (1985). Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1):100–132.

Haritaoglu, I., Harwood, D., and Davis, L. S. (1998). W 4 s: A real-time system for detecting and tracking people in 2 1/2d. In *European Conference on computer vision*, pages 877–892. Springer.

Hasinoff, S. W. and Kutulakos, K. N. (2007). A layer-based restoration framework for variable-aperture photography. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Howard, I. P., Rogers, B. J., et al. (1995). *Binocular vision and stereopsis*. Oxford University Press, USA.

Huang, T. (1996). Computer vision: Evolution and promise.

Ikoma, H., Nguyen, C. M., Metzler, C. A., Peng, Y., and Wetzstein, G. (2021). Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE.

Kanade, T. and Okutomi, M. (1994). A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE transactions on pattern analysis and machine intelligence*, 16(9):920–932.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., and Bry, A. (2017). End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75.

Klette, R. (2014). *Concise computer vision*. Springer.

Levin, A., Fergus, R., Durand, F., and Freeman, W. T. (2007). Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70–es.

Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., and Zhang, J. (2018). Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820.

Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., and Kautz, J. (2017). Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30.

Luo, W., Schwing, A. G., and Urtasun, R. (2016). Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estima-

tion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048.

Metzler, C. A., Ikoma, H., Peng, Y., and Wetzstein, G. (2020). Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385.

Morris, T. (2004). *Computer vision and image processing*. Palgrave Macmillan.

Murray, D. and Little, J. J. (2000). Using real-time stereo vision for mobile robot navigation. *autonomous robots*, 8(2):161–171.

Okutomi, M., Katayama, Y., and Oka, S. (2002). A simple stereo algorithm to recover precise object boundaries and smooth surfaces. *International Journal of Computer Vision*, 47(1):261–273.

Oâ€™Riordan, A., Newe, T., Dooly, G., and Toal, D. (2018). Stereo vision sensing: Review of existing systems. In *2018 12th International Conference on Sensing Technology (ICST)*, pages 178–184. IEEE.

Pang, J., Sun, W., Ren, J. S., Yang, C., and Yan, Q. (2017). Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895.

Paramonov, V., Panchenko, I., Bucha, V., Drogolyub, A., and Zagoruyko, S. (2016). Depth camera

based on color-coded aperture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9.

Pentland, A. P. (1987). A new sense for depth of field. *IEEE transactions on pattern analysis and machine intelligence*, (4):523–531.

Praveen, S. (2020). Efficient depth estimation using sparse stereo-vision with other perception techniques. *Coding Theory*, page 111.

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.

Shamsafar, F., Woerz, S., Rahim, R., and Zell, A. (2022). Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 2417–2426.

Short, N. J. (2009). *3-D Point Cloud Generation from Rigid and Flexible Stereo Vision Systems*. PhD thesis, Virginia Tech.

Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., and Wetzstein, G. (2018). End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):1–13.

Song, X., Zhao, X., Hu, H., and Fang, L. (2018). Edgestereo: A context integrated residual pyramid network for stereo matching. In *Asian Conference on Computer Vision*, pages 20–35. Springer.

Sonka, M., Hlavac, V., and Boyle, R. (2014). *Image processing, analysis, and machine vision*. Nelson Education.

Tan, S., Wu, Y., Yu, S.-I., and Veeraraghavan, A. (2021). Codedstereo: Learned phase masks for large depth-of-field stereo. *arXiv preprint arXiv:2104.04641*.

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.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426.

Umbaugh, S. E. (2010). *Digital image processing and analysis: human and computer vision applications with CVIPtools*. CRC press.

Wang, C., Sahin, E., Suominen, O., and Gotchev, A. (2014). Depth estimation by combining stereo matching and coded aperture. In *2014 IEEE Visual Communications and Image Processing Conference*, pages 291–294. IEEE.

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.

Xu, H. and Zhang, J. (2020). Aanet: Adaptive aggregation network for efficient stereo matching. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1959–1968.

Yan, Z. and Zha, H. (2019). Flow-based slam: From geometry computation to learning. *Virtual Reality & Intelligent Hardware*, 1(5):435–460.

Yang, G., Zhao, H., Shi, J., Deng, Z., and Jia, J. (2018). Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 636–651.

Zbontar, J. and LeCun, Y. (2015). Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599.

Zhang, F., Prisacariu, V., Yang, R., and Torr, P. H. (2019). Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194.

Zhang, J., Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., and Quan, L. (2021). Learning stereo matchability in disparity regression networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1611–1618. IEEE.

Zhou, C., Lin, S., and Nayar, S. K. (2011). Coded aperture pairs for depth from defocus and defocus deblurring. *International journal of computer vision*, 93(1):53–72.

Zhou, K., Meng, X., and Cheng, B. (2020). Review of stereo matching algorithms based on deep learning. *Computational intelligence and neuroscience*, 2020.