Diseño y Modelado Matemático de un Sistema Optoelectrónico Multisensor para Adquisición de Información Espectral y de Profundidad

Elizabeth Juliana Martínez Ayala

Juan Pablo Cuadrado Flechas

Sergio Andrés Urrea Vecino

Trabajo de Grado para optar al título de Ingeniero Electrónico

Director:

Ph.D(c) Hans Yecid Garcia Arenas

Codirectores:

Ph.D Said Pertuz Arroyo

Ph.D Henry Arguello Fuentes

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2022

Dedicatoria

Al grupo de investigación en diseño de algoritmos y procesamiento de datos multidimensionales HDSP.

Agradecimientos

Agradecimientos al Grupo de Investigación en Diseño de Algoritmos y Procesamiento de Datos Multidimensionales HDSP.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Resumen

Título: Diseño y modelado matemático de un sistema optoelectrónico multisensor para adquisición de información espectral y de profundidad [*]

Autores: Elizabeth Juliana Martínez Ayala, Juan Pablo Cuadrado Flechas, Sergio Andrés Urrea Vecino.

Palabras claves: Estéreo visión, LiDAR, Profundidad, Desenfoque, Fusión, Aprendizaje Profundo.

Descripción: Una imagen es la representación visual de una escena, que tradicionalmente es adquirida por uno o varios instrumentos optoelectrónicos que capturan información de las distintas dimensiones que la conforman. Por ello actualmente diversas plataformas han propuesto arquitecturas que permiten capturar información de distintas dimensiones, las cuales están compuestas por cámaras, RGB, de infrarrojo y múltiples sensores como LiDAR, que permiten desarrollar aplicaciones diversas, entre las que se encuentra la conducción autónoma, la realidad virtual y la reconstrucción de escenas en tres dimensiones (3D). Específicamente en este trabajo se propone una arquitectura de adquisición de información de espacio, profundidad y color que está compuesta por dos cámaras RGB, y una cámara de infrarrojo cercano. Además se desarrolló una red de inteligencia artificial, red DNet, que permitió la reconstrucción de mapas de profundidad, y la estimación de la mejor distancia entre cámaras RGB para imágenes sintéticas e imágenes de implementación. Específicamente, para la evaluación de la red empleando imágenes sintéticas de entrenamiento, se obtuvo que siguiendo las métricas de PSNR y SSIM la mejor configuración para estimar un mapa de profundidad es de un metro de distancia entre cámaras y un nivel de profundidad de 1 metro. Por otra parte, en validación se encontró que se tienen buenas reconstrucciones para distintos de niveles de profundidad, si se conserva en cada caso una distancia entre cámaras de 1 metro. Finalmente, se encuentra que la red tiene un rendimiento aceptable para imágenes capturadas en entorno real no controlado con el sistema implementado.

---

[*] Trabajo de grado

Abstract

Title: Design and mathematical modeling of a multisensor optoelectronic system for acquisition of spectral and depth information [*]

Author: Elizabeth Juliana Martinez Ayala, Juan Pablo Cuadrado Flechas, Sergio Andrés Urrea Vecino [**]

Keywords: Stereo Vision, LiDAR, Depth, Defocus, Fusion, Deep Learning

Description: An image is the visual representation of a scene, which has been traditionally acquired by one or more optoelectronic instruments that capture information from different dimensions that describes a scene. For this reason, nowadays, diverse platforms had proposed architectures that capture the information of the dimensions that compose a scene, these platforms are usually built using a set of RGB, infrared cameras, and multiple sensors such as Lidar. Specifically, these platforms allow the development of diverse applications, like autonomous driving, virtual reality, and the reconstruction of scenes in three dimensions (3D). Specifically, this work proposed an architecture for the acquisition of space, depth, and color information, composed by two RGB cameras and a near-infrared camera. In addition, an artificial intelligence network, DNet network, was developed, to estimate depth map reconstructions, and to estimate the performance of the model varying the distance between RGB cameras for synthetic and implementation images. Specifically, for the evaluation of the network using synthetic training images, it was found that following the PSNR and SSIM metrics, the best distance between the two cameras to estimate a depth map is a distance of one meter for a depth level of one meter. For the validation process, it was found that if the two cameras were placed at a distance of one meter the reconstruction is preserved for each depth level. Finally using the implementation of the system proposed, it is found that the network has acceptable performance for images captured in a real uncontrolled

---

[*] Bachelor Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Ph.D(c) Hans Yecid Garcia Arenaso, Doctorado en Ingeniería eslectrónica.

environment.

Introducción

Una imagen es la representación visual de una escena que tradicionalmente se adquiere utilizando instrumentos optoelectrónicos, como las cámaras digitales, las cuales permiten capturar la información de la escena y representarla discretamente como una matriz bidimensional de muestras discretas. Sin embargo, se conoce que una escena está compuesta por múltiples dimensiones que pueden ser descritas como la propagación de una onda reflejada desde la escena, en donde generalmente esta onda tiene componentes de tiempo t, espacio (x,y) , de amplitud A, longitud de onda λ, de polarización, de ángulo α y de fase Φ las cuales pueden ser capturada utilizando diversos sensores optoelectrónicos o arquitecturas ópticas como: las cámaras de vídeo, las cámaras hiper espectrales, cámaras infrarrojas, sensores de profundidad como: el sensor LiDAR, la arquitectura de estero visión, y los elementos ópticos difractivos. Actualmente, se han desarrollado diversos sistemas compuesto por múltiples sensores, como el sistema NYU compuesto por un sensor RGB y un sensor de profundidad [Silberman et al., 2012], [Ma and Karaman, 2018]; la plataforma Annieway de conducción autónoma que cuenta con un sistema de sensores de estéreo visión, flujo óptico, odometría visual[Geiger et al., 2013], [Fritsch et al., 2013], entre otras. Donde estas plataformas son utilizadas para el desarrollo de aplicaciones basadas en inteligencia artificial que permiten realizar tareas de predicción[Chen and Blum, 2009], clasificación [Amarsaikhan et al., 2012], segmentación [Du and Gao, 2017] y reconstrucción[Kennedy et al., 2007]. Algunas de estas aplicaciones se desarrollan en campos de agricultura, **NIR** [Camacho et al., 2018]; desarrollo de vehículos autónomos, detección y seguimiento de objetos 3D. Específicamente en este trabajo se propone una plata-

forma de adquisición de información de espacio, profundidad y color compuesta por dos cámaras *RGB*, y una cámara de infrarrojo cercano *NIR*. Para esto se diseña e implementa un algoritmo de inteligencia artificial que permita evaluar el desempeño del parámetro de distancia entre cámaras para la reconstrucción de los mapas de profundidad. Todo lo anterior, con el fin de poder capturar y fusionar la información obtenida por cada sensor para una reconstrucción de un cubo de datos que contenga información de profundidad e información espectral .

1. Objetivos

Objetivo general

Diseñar y modelar matemáticamente un sistema optoelectrónico multisensor para adquisición de información espectral y de profundidad, teniendo en cuenta las restricciones de entornos interiores y exteriores.

Objetivos específicos

+ Modelar matemáticamente el proceso de adquisición de una arquitectura de medición de profundidad, una de infrarrojo cercano y una cámara RGB.

+ Diseñar los parámetros libres para la arquitectura de adquisición de profundidad, información RGB y del infrarrojo cercano.

+ Diseñar e implementar un algoritmo de fusión que incluya la obtención de una imagen espectral del rango visible e infrarrojo cercano y su fusión con información de profundidad.

+ Evaluar en simulación y en implementación el rendimiento de un sistema optoelectrónico empleando el algoritmo de fusión desarrollado para los parámetros libres diseñados.

2. Estado del arte

2.1. Arquitecturas de adquisición de profundidad

La estimación de profundidad es una tecnología crucial para distintas aplicaciones como: robótica [Weiss and Biber, 2011], la conducción autónoma [Wang et al., 2019] y la realidad aumentada [Li et al., 2019]. Esto ha conllevado al desarrollo de diversos sistemas de medición de profundidad tales como estéreo visión [Park et al., 2018]; LiDAR [Tsai et al., 1998]; profundidad a partir del desenfoque [Weihua and Shuang, 2019]. Sin embargo, se conoce que dichos sistemas pueden reducir su precisión al ser sometidos a diversas condiciones del entorno, como lo son: cambios en la iluminación, el tipo de sensor implementado, los parámetros implementados en el diseño de las arquitecturas de profundidad previamente mencionadas como la distancia entre cámaras, resolución de los sensores, el campo de visión de los sensores, entre otros. Lo que hace necesario estudiar las diferentes técnicas de adquisición de profundidad con el fin determinar cual de ellas resulta ser la más adecuada frente a los ambientes. En esta sección se presentará la definición y la comparativa de tres arquitecturas o modelos de adquisición de profundidad con el fin de evaluar y elegir el sensor con el que se realizará el modelado, las pruebas de simulación y la implementación de la arquitectura de fusión. Específicamente, se comparan los sistema de estéreo visión, sensor LiDAR y arquitectura de profundidad por desenfoque teniendo en cuenta parámetros como: tipo sensor, iluminación del entorno de trabajo, limitaciones ópticas, y costo, estas comparaciones van a ser presentadas de manera extendida en la sección Comparativa de arquitecturas de profundidad.

2.2. Estéreo visión

Es una técnica de adquisición de imágenes que puede proporcionar mediciones de profundidad de los objetos de una escena respecto a dos sensores. El principio matemático de la estéreo visión es similar a la percepción 3D en la visión humana y se fundamenta en la triangulación de rayos desde múltiples puntos de vista [Zou and Li, 2010]. El modelo de estéreo visión presentado está compuesto por dos sensores pasivos, usualmente de la misma referencia, separados a una distancia β, los cuales realizan las capturas de una escena de manera simultánea como se muestra en la figura 2.1.

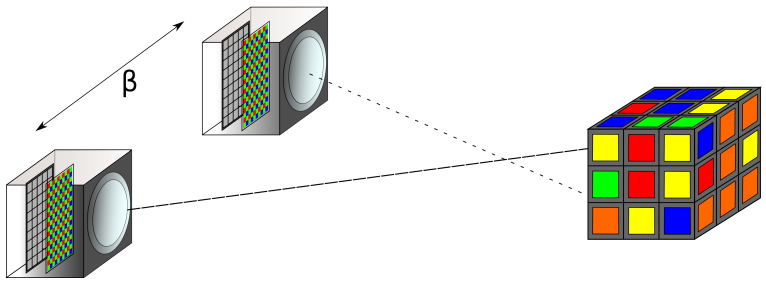

Figure 2.1. Ilustración de un modelo básico de estéreo visión compuesto por dos cámaras separadas a una distancia β.

No obstante, las imágenes capturadas por los sensores requieren ser procesadas por algoritmos para la estimación de la profundidad de la escena. Este tratamiento parte del concepto de la disparidad que se define como la diferencia entre píxeles de un objeto A en la imagen capturada en el sensor derecho respecto a el mismo objeto A en la imagen capturada por el sensor izquierdo como se observa en la figura 2.2.

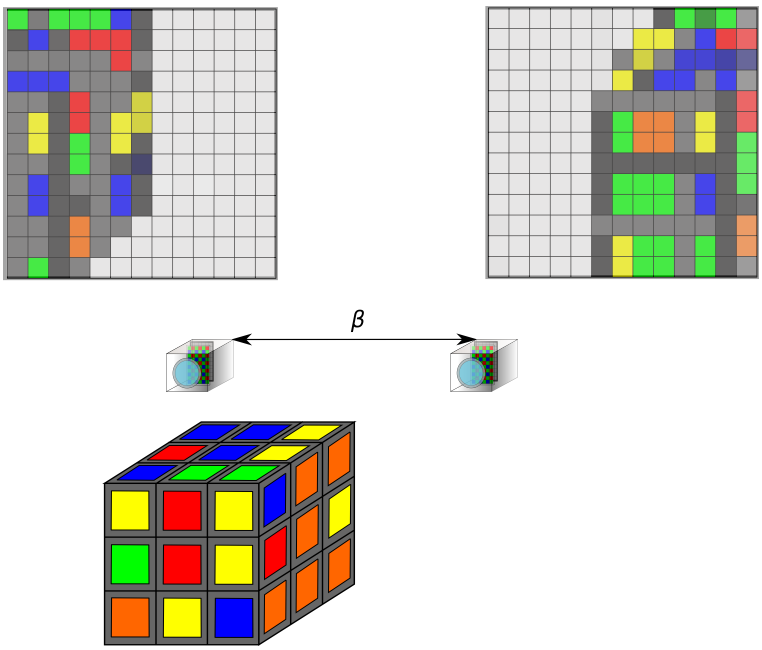

Figure 2.2. Ilustración de la adquisición de la disparidad a partir de las imágenes RGB capturadas en el sensor Izquierdo y derecho.

Donde luego es posible encontrar los mapas de profundidad de la escena a partir de la siguiente expresión matemática:

$$\mathbf{Z} = \frac{f\beta}{\mathbf{D}}, \tag{2.1}$$

sabiendo que **D** es el mapa de disparidad, **Z** es el mapa de profundidad, β es la distancia entre cámaras y f es la distancia focal del lente de los sensores, en la figura 2.3 se presenta una ilustración de la relación entre un mapa de profundidad y un mapa de disparidad.

Y el mapa de disparidad se puede aproximar sabiendo que

$$d = \hat{x_r} - \hat{x_L} = D\frac{uF - \bar{v}u + F\bar{v}}{Fu} = Df(u) \tag{2.2}$$

Donde $x_r = (x_0, y_0)$ y $x_L = (x_0 + D, y_0)$, donde D varía según las coordenadas (x, y, z), haciendo que el corrimiento se defina como una función no lineal así:

$$x_L = (x_0 + D(x, y, z), y_0) \tag{2.3}$$

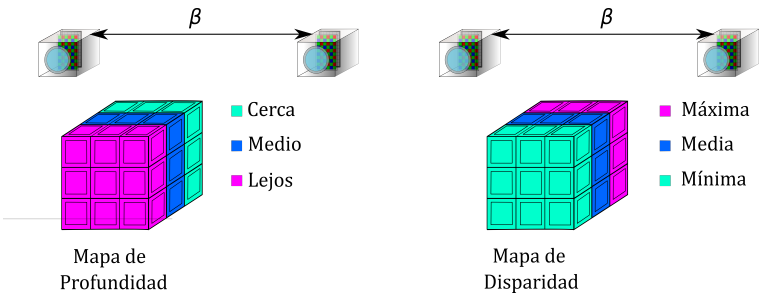

Figure 2.3. Ilustración del un mapa de profundidad y un mapa de disparidad.

2.3. Detección de luz y alcance LiDAR

LiDAR es un sensor optoelectrónico activo de detección y rango de luz que permite crear una representación 3D de una escena. La forma en la que un sistema LiDAR permite crear una escena 3D parte de que el emisor de láser del sistema emite rayos de luz, monocromática y coherente [Polyakov et al., 2014] , en forma de pulsos de ondas que viajan a través del espacio hasta encontrar un objeto que refleje la luz hacia el receptor. Lo anterior es posible dado que la velocidad de la luz en el aire, medio que se considera para la propagación de la luz, es aproximadamente constante. La luz láser reflejada por el objeto es detectada y analizada por los receptores en el sensor LiDAR. Para calcular la distancia entre el sensor y un objeto, los receptores registran el tiempo que le toma

al pulso láser desde que es enviado hasta que es recibido por el sensor así:

$$distancia = \frac{tiempo * velocidaddelaluz}{2} \tag{2.4}$$

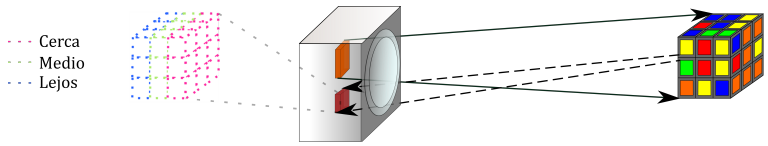

Figure 2.4. En esta imagen se presenta la formación de la nube de puntos que se obtiene después de que el receptor atrape los rayos de luz que vienen del emisor. Fuente: Propia.

Específicamente, para las imágenes 3D o mapas de profundidad, el receptor de LiDAR captura la escena como una nube de puntos que representa los valores de profundidad de cada elemento de una escena. Luego esta nube de puntos es evaluada por algoritmos que reconstruyen los mapas de profundidad teniendo en cuenta la información adquirida, la cercanía espacial de los puntos y la información a priori de la escena.

2.4. Profundidad por desenfoque

La profundidad a partir del desenfoque es una técnica que permite estimar la profundidad a partir de la toma de una única o múltiples capturas. [Schechner and Kiryati, 2000]

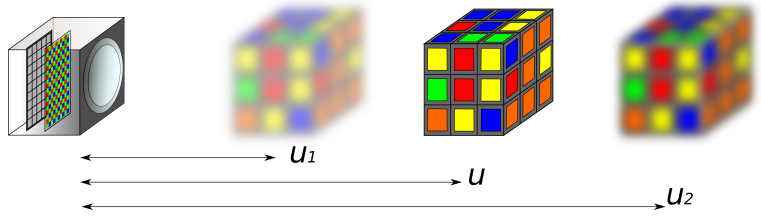

Figure 2.5. En esta imagen se presenta el ejemplo de tres objetos, dónde el objeto a la distancia u y los objetos u_1 y u_2 se encuentran en desenfoque. Fuente: Propia

Para estimar la profundidad de un objeto en una imagen a partir de las medidas de desenfoque, es necesario definir la geometría de la Figura 2.6 que describe el camino de los rayos de luz de dos objetos O_1 y O_2, que pasan a través de un lente de radio r, que está a una distancia u_1 y u respectivamente.

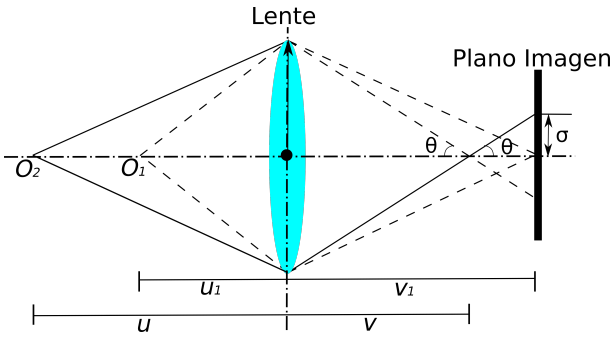

Figure 2.6. En esta imagen se presenta la geometría y trazado de rayos de dos objetos O_1 en foco, y O_2 en desenfoque. Fuente: Propia.

Se observa entonces, que los rayos de luz de los objetos O_1 y O_2 convergen a diferentes distancias v_1 y v, notando que el objeto O_1 se encuentra en foco y los rayos de luz que provienen del objeto O_2 tienen un desenfoque de radio σ definido como

$$\sigma = \frac{r(v_1 - v)}{v} \tag{2.5}$$

entonces, para calcular la profundidad a partir del desenfoque primero se define la ecuación de lentes para de ahí despejar la distancia a la que un objeto se encuentra en foco u así,

$$\frac{1}{f} = \frac{1}{u} + \frac{1}{v} \tag{2.6}$$

ahora despejando u de 2.6 se tiene que,

$$u = \frac{fv}{v-f} \tag{2.7}$$

ahora teniendo en cuenta que la distancia focal f es una distancia fija y definiendo a $v = v_1$ y $u = u1$, se podría decir que el objeto O_2 está enfocado en

$$u_1 = \frac{fv_1}{v-f_1} \tag{2.8}$$

Entonces se puede decir que para un objeto O_1 donde $u > u_1$ y que está desenfocada con un radio de desenfoque σ, tomando la profundidad como $D = u$, e igualando las ecuaciones 2.5 y 2.8, se tiene que la profundidad D de un objeto está definida por:

$$D = \frac{frv_1}{rv_1 - f(r-\sigma)} \tag{2.9}$$

3. Modelos de adquisición

3.1. Cámaras RGB.

Una cámara RGB contiene un sensor optoelectrónico que permite capturar una escena en la representación del espacio de color RGB. En el proceso de captura, se parte de que el sensor contiene 3 filtros de color excluyentes y complementarios que filtran las longitudes de onda correspondientes a los canales de color R, G y B, que siguen el patrón Bayer como se muestra en la figura 3.2, debido a que al aplicar estos filtros, se resulta con muchos píxeles sin información en cada banda de color, la cámara aplica un proceso llamado *demosaicing* [Malvar et al., 2004] , el cual interpola los píxeles con información de la banda de color para estimar el valor de los demás píxeles. Sin embargo, en este trabajo se asume un píxel como el resultante del proceso para un cuadrado de 2×2 píxeles como se muestra en la figura 3.1. Matemáticamente el proceso de adquisición de la

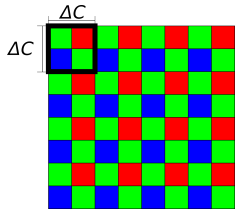

Figure 3.1. Mosaico de filtros correspondiente al patrón de Bayer.Fuente: Propia.

imagen esta definido por la ecuación

$$\mathbf{P}_{i,j,l} = \int_{\lambda_2}^{\lambda_1} \int_{(j-1)\Delta c}^{j\Delta c} \int_{(i-1)\Delta c}^{i\Delta c} \Gamma_l(\lambda) F(x,y,\lambda)\, dx\, dy\, d\lambda, \tag{3.1}$$

Donde $F(x,y,\lambda)$ representa la escena antes de ser discretizada, $\Gamma(\lambda)$ representa la respuesta espectral de la cámara, que se define como la fracción de la intensidad de luz que es capaz de absorber el detector por cada longitud de onda $\lambda_1 = 400[nm]$, $\lambda_2 = 750[nm]$ y Δc el tamaño del píxel.

Figure 3.2. Esquema de adquisición de una cámara RGB.Fuente: Propia.

3.2. Cámara Infrarroja

Es un sensor optoelectrónico que captura una escena en el rango del espectro electromagnético del infrarrojo. Los sensores infrarrojos pueden ser sensores: pasivos cuando solo están formados por un foto transistor o activos cuando su configuración se basa en la combinación de un emisor y un receptor. Además de esta división, los sensores de infrarrojo se subdividen según el rango del espectro para el cual son diseñados [Capper and Elliott, 2013], en infrarrojo cercano(NIR), infrarrojo de longitud de onda corta(SWIR), infrarrojo de onda media(MWIR), e infrarrojo de onda larga(LWIR), a manera de ilustración en la Figura 3.3 se observa el rango para cada una de la divisiones en el espectro del infrarrojo con el respectivo material del que son creados los sensores. Por tal motivo, en el mercado se presentan diversos sensores de infrarrojo que capturan rangos determinados. De la Tabla 3.1 podemos observar distintos sensores comerciales, sus rangos de captura, su precio y resolución espectral para cámaras de infrarrojo y comerciales.

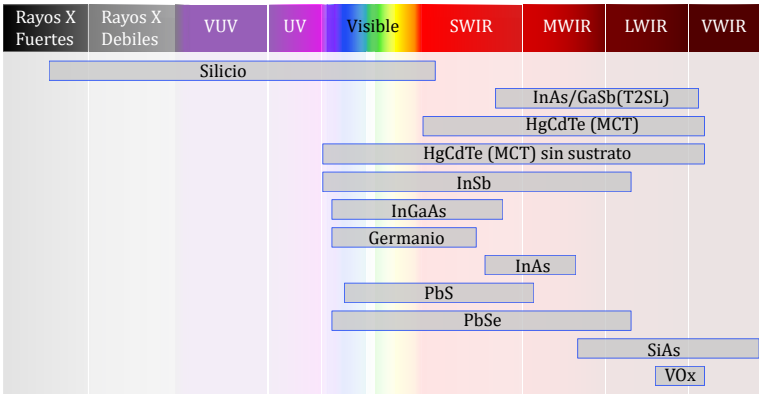

Figure 3.3. Subdivisión de los materiales de construcción de los sensores infrarrojos según el espectro electromagnético

Sensor	Referencia	Resolución	Rango espectral	Precio (USD)
IR	Seek Thermal - Shotpro	320 x 240	7.5 – 14 [μm]	712.25
	XGUANMETER	120 x 160	8 – 14 [μm]	356.64
	FLIR ONE Pro	80x60	8 – 14 [μm]	234.35
RGB	LUMIX DMC-FZ300K	3840 x 2160	Visible	456.47
	LUMIX DMC-FZ70	3840 x 2160	Visible	303.63
	Minolta MN35Z	920 x 1080	Visible	267.13

Tabla 3.1. Comparación de resolución y precio de las cámaras comerciales de infrarrojo y RGB.

De esta tabla cabe resaltar que la resolución del sensor tiene una relación inversa con el número de bandas infrarrojas que puede capturar y que también tiene una relación directa con el precio de la misma. Así mismo, se observa también que comparada con los sensores *RGB*, las cámaras de infrarrojo presentan baja resolución espacial y un elevado costo

$$\mathbf{G}_{i,j} = \int_{\lambda_3}^{\lambda_4} \int_{(j-1)\Delta m}^{j\Delta m} \int_{(i-1)\Delta m}^{i\Delta m} \Gamma_{IR}(\lambda) F(x,y,\lambda)\, dx\, dy\, d\lambda, \tag{3.2}$$

donde λ_3 representan la longitud de onda del limite inferior del ancho de banda del espectro del infrarrojo detectado por el sensor, λ_4 la longitud de onda del limite superior del ancho de banda

del espectro del infrarrojo detectado por el sensor, $\Gamma_{IR}(\lambda)$ representa la respuesta espectral de la cámara, y Δm el tamaño del píxel del sensor del infrarrojo.

3.3. Comparativa de arquitecturas de profundidad

Como se describió en las subsecciones , existen distintas formas de obtener o calcular la profundidad de un objeto a partir de la información que recolectan los sensores optoelectrónicos que las componen. Entre las arquitecturas evaluadas se encuentran la arquitectura LiDAR, la arquitectura de estéreo visión y la técnica de estimación de profundidad a partir de desenfoque. A continuación se presenta una comparativa de las ventajas y desventajas que tiene cada uno de los sistemas con el fin de evaluar uno de ellos en implementación teniendo en cuenta variables como lo son, complejidad del sistema, facilidad de calibración, disponibilidad del sensor, resolución del sensor y el costo del mismo.

Comenzando la comparativa, se presenta la arquitectura más costosa, que es, la arquitectura LiDAR. Esta arquitectura tiene como principal ventaja su alta precisión, captura de datos a alta velocidad, al rededor de 20 revoluciones por segundo, además presenta baja sensibilidad ante las perturbaciones de la temperatura, lo cual permite tener un sistema calibrado. Sin embargo, LiDAR presenta una alta sensibilidad a los diversos niveles de reflectividad y a la variación de las condiciones climáticas como la lluvia, la niebla o el polvo. Por otra parte, LiDAR se presenta como la arquitectura de profundidad evaluada más costosa teniendo un precio cercano a los 580 USD para sus versiones más económicas, lo que hace necesario realizar la evaluación de otros sistemas de adquisición más asequibles y para ellos se evalúan la el sistema de estéreo visión o la técnica de profundidad a partir de desenfoque. Para esta última, se tiene como ventaja principal que es

un sistema óptico con un montaje muy compacto, ya que está formado a partir de un único lente. Lo anterior, permite evitar aberraciones como la oclusión debido a que el lente objetivo siempre está viendo la misma escena. Sin embargo, al tener una arquitectura monocular para estimación de profundidad, se requiere la toma de múltiples capturas lo cual aumenta significativamente el tiempo de adquisición de datos y de reconstrucción del mapa de profundidad. Por esto, se plantea la evaluación del ultimo sistema, que es un sistema bifocal de estéreo visión, que si bien posee una arquitectura de mayor tamaño que es sensible a las perturbaciones externas y a aberraciones por oclusión, tiene ventajas como que es un sistema de fácil adquisición que permite la captura de imágenes a alta resolución en un tiempo corto de adquisición. Por tal motivo después de realizar estas comparativas se decide elegir el sistema de estéreo visión como el sistema que va a ser implementado en simulación debido a que es el sistema que permite tener imágenes de más alta resolución a un menor costo, y con menor tiempo de adquisición. En la tabla 3.2 se presentan otros datos relevantes a la hora de escoger la arquitectura para adquisición de mapas de profundidad.

	Tiempo de captura	**Forma de captura**	**Tipo de sensor**
Profundidad Mediante Desenfoque	τ	Estimación	Pasivo
Estéreo Visión	τ	Estimación	Pasivo
LiDAR	$\frac{M \times N}{T}$	Adquisición	Activo

Tabla 3.2. Comparativa arquitecturas de adquisición de profundidad, donde τ es el tiempo de integración de las cámaras, $M \times N$ el tamaño en píxeles de la escena a adquirir con el sensor LiDAR, y T la cantidad de puntos que puede adquirir el mismo sensor en un segundo, típicamente τ es mucho menor a $\frac{M \times N}{T}$ dado el tamaño de las escenas a capturar.

Para reducir el tiempo de captura de la escena, los sensores LiDAR capturan nubes de puntos espaciados con los que reconstruyen el mapa de profundidad.

3.4. Algoritmos de fusión basados en inteligencia artificial

Los algoritmos de fusión han sido ampliamente utilizados en el estado del arte para dar solución a problemas que utilicen información de mas de una fuente para generar una salida aprovechando lo mejor las múltiples fuentes, como puede ser la súper resolución. Un ejemplo puede ser el uso de nubes de puntos provenientes de LiDAR e imágenes RGB con el objetivo de generar mapas de profundidad de alta resolución [Li et al., 2013]. Otro ejemplo de la aplicación es la fusión de imágenes multiespectrales (alta resolución espacial y baja resolución espectral), con imágenes hiper espectrales (alta resolución espectral y baja resolución espacial) para crear imágenes de alta resolución espacial y espectral [Vargas et al., 2019] [Ramirez and Arguello, 2019]. Además estos algoritmos son usados en la fusión de imágenes con información visual e información del infrarrojo [Liao et al., 2020] [Wang et al., 2020]. Específicamente, para la solución del problema de fusión a partir de algoritmos basados en inteligencia artificial, en el estado del arte se presentan diferentes arquitecturas como por ejemplo lo son las redes *GAN* [Liao et al., 2020], que son redes neuronales compuestas por un generador, un discriminador y una métrica de control. Otro ejemplo de arquitectura de red usada para resolver problemas de fusión es la red convolucional profunda *VGG* [Wang et al., 2020] [Fu et al., 2019] desarrollada por *Visual Geometry Group* de la Universidad de Oxford.

Otro enfoque del uso de inteligencia artificial en los algoritmos de fusión, es el uso de múltiples redes neuronales para la extracción de características y fusión, como un ejemplo de esto, en [Lakshmipriya et al., 2020] se utilizan redes neuronales para cada uno de los canales de color con

el fin de fusionar dos imágenes capturadas con diferentes sensores, variando parámetros como iluminación, enfoque, perspectivas, de la misma escena, para obtener como resultado una imagen libre de desenfoque,específicamente, este artículo se hacen pruebas con 3 arquitecturas diferentes las cuales son: Alexnet, que esta compuesta por 5 capas convoluciones y 3 totalmente conectadas, VGG-16, contando con 16 capas aprendibles y una distribución de capas similar a la de Alexnet, pero remplaza filtros con kernel de gran tamaño por secuencias de filtrado con kernels de tamaño 3x3 , y GoogLeNet, que combina capas convolucionales con filtros que extraen características a diferentes resoluciones. Otro ejemplo es el uso de múltiples arquitecturas de inteligencia artificial al mismo tiempo dentro de una sola arquitectura, un ejemplo de este enfoque es presentado en [Lavinia et al., 2016].

4. Problemas Inversos

Un problema inverso consiste en estimar la información de una escena x a partir de la información que arroja un conjunto de medidas u observaciones y [Groetsch and Groetsch, 1993]. Los problemas inversos son problemas matemáticos que nos permiten encontrar parámetros que no podemos observar o sensar directamente y pueden ser descritos de la siguiente manera

$$y = H(x) \tag{4.1}$$

donde H es un operador lineal que describe la relación entre los datos y y la información del modelo x que es la que se desea encontrar. Específicamente, en el caso de un problema inverso

lineal discreto, lo anterior puede ser reescrito como

$$\mathbf{y} = \mathbf{H}\mathbf{x} \tag{4.2}$$

donde $\mathbf{H}$ es una matriz que modela el sistema de adquisición, $\mathbf{y}$ es el vector de observaciones, y $\mathbf{x}$ es el vector de parámetros a encontrar. Específicamente, la solución al sistema optoelectrónico multisensor podría realizarse definiendo cada sistema a partir de sus operadores lineales $\hat{\mathbf{H}}$ para el sensor RGB, $\bar{\mathbf{H}}$para el sensor NIR, e idealmente $\dot{\mathbf{H}}$ para la arquitectura de profundidad, como

$$[\hat{\mathbf{y}}, \bar{\mathbf{y}}, \dot{\mathbf{y}}]^T = [\hat{\mathbf{H}}, \bar{\mathbf{H}}, \dot{\mathbf{H}}]^T \mathbf{x}, \tag{4.3}$$

donde el vector solución $\mathbf{x}$ se obtiene resolviendo el problema de optimización

$$\underset{\mathbf{x}}{\operatorname{argmin}} ||\hat{\mathbf{H}}\mathbf{x} - \hat{\mathbf{y}}||_2 + ||\bar{\mathbf{H}}\mathbf{x} - \bar{\mathbf{y}}||_2 + ||\dot{\mathbf{H}}\mathbf{x} - \dot{y}||_2 \tag{4.4}$$

Pero, como se definió en la sección el modelo de estéreo visión no es un problema lineal, por lo que se decide plantear el problema como

$$\underset{\mathbf{x}}{\operatorname{argmin}} \|\mathfrak{M}\{\mathbf{y}\} - \mathbf{x}\|_2 \tag{4.5}$$

donde ,$\mathfrak{M}$ es una red de inteligencia artificial, que puede resolver el problema entregando el vector solución $\mathbf{x}$, y que además permite realizar la evaluación de los parámetros libres ,como la distancia

entre cámaras, que definen la arquitectura, y que son utilizados como herramienta para optimizar el sistema optoelectrónico.

5. Propuesta: Red DNet

Durante esta sección se presenta la propuesta de una arquitectura de inteligencia artificial que tiene como objetivo principal la reconstrucción del mapa de profundidad de una escena, además servirá para evaluar los mejores parámetros ópticos de un sistema de múltisensor para la adquisición de información espectral y de profundidad, con el fin de facilitar la lectura de la sección se presenta la nomenclatura de la variables implementadas en la Tabla 5.1

Nomenclatura	Nombre	Tamaño
$\mathbf{S}$	Imagen Sensor Derecho	$256 \times 256 \times 3$
$\overline{\mathbf{S}}$	Imagen Sensor Izquierdo	$256 \times 256 \times 3$
$\mathbf{G}$	Imagen Infrarrojo cercano	$64 \times 64 \times 1$
$\mathbf{D}$	Mapa de disparidad	$256 \times 256 \times 1$
β	Distancia entre cámaras	Escalar
$\mathbf{Z}$	Mapa de profundidad	$256 \times 256 \times 1$
$\mathbf{V}$	Imagen RGB-NIR sensor derecho	$256 \times 256 \times 4$

Tabla 5.1. Tabla de nomenclatura de los parámetros de la Red D-Net

5.1. Bloques propuestos

Como se mencionó anteriormente se desea encontrar las distancias óptimas a las que deben de estar situados dos sensores *RGB*, que hacen parte de sistema optoelectrónico multisensorial, con el fin de obtener un mapa de profundidad de alta resolución espacial. Para ello, en este trabajo se propone la red DNet, que es una red compuesta por 2 bloques principales: Bloque de profundidad que se encarga de generar el mapa profundidad $\mathbf{Z}$ a partir de la relación de las capturas adquiridas posicionando los sensores a una distancia β,y bloque de fusión que se enfoca en la fusión de datos

que presenten distintas características espaciales y espectrales como se ilustra en la Figura. 5.1. Como primer paso, se tiene que a la DNet ingresan tres imágenes cuadradas: Imagen del Sensor Derecho **S**, Imagen del Sensor Izquierdo $\overline{\mathbf{S}}$ y una Imagen de Infrarrojo cercano **G**. Específicamente, $\overline{\mathbf{S}}$ y **S** ingresan al bloque de profundidad donde se estima un mapa de profundidad **Z** . De manera paralela, **G** y **S** ingresan al bloque de fusión que da como salida una imagen VNIR **V**. Cabe resaltar, que para este proyecto se plantea una optimización basada en aprendizaje profundo, por lo cual se requiere imágenes de infrarrojo de tamaño similar a las cámaras del estado del arte, debido a que durante la búsqueda solo se encuentran conjuntos de datos de adquisición de información a partir múltiples sensores con ausencia de información en el infrarrojo [Yang et al., 2019][Martull et al., 2012] [Menze and Geiger, 2015]. Por lo cual se asume un submuestreo de las imágenes RGB en escala de grises teniendo en cuenta la relación entre tamaños de los sensores de las cámaras RGB y las cámaras infrarrojas.

5.1.1. Bloque Profundidad. El bloque de profundidad se encarga de generar el mapa de profundidad a partir de la imagen derecha **S** e izquierda $\overline{\mathbf{S}}$ capturadas por los sensores. Este bloque de profundidad presentado en la figura 5.2 tiene como entradas la imagen capturada por el sensor derecho **S** y la imagen capturada por el sensor izquierdo $\overline{\mathbf{S}}$, donde las entradas ingresan al bloque de extracción de características S-Z-1 y S-Z-2 respectivamente. En estructura estos bloques son idénticos, pero no comparten sus pesos. Los bloques S-Z se encargan de la extracción de características de la imagen de entrada con el fin de obtener una aproximación a un mapa de profundidad. Estos bloques están basados en la arquitectura U-net descrita en la Figura 5.3 donde al inicio se realizan operaciones de reducción de resolución utilizando convoluciones o bloques Max Poo-

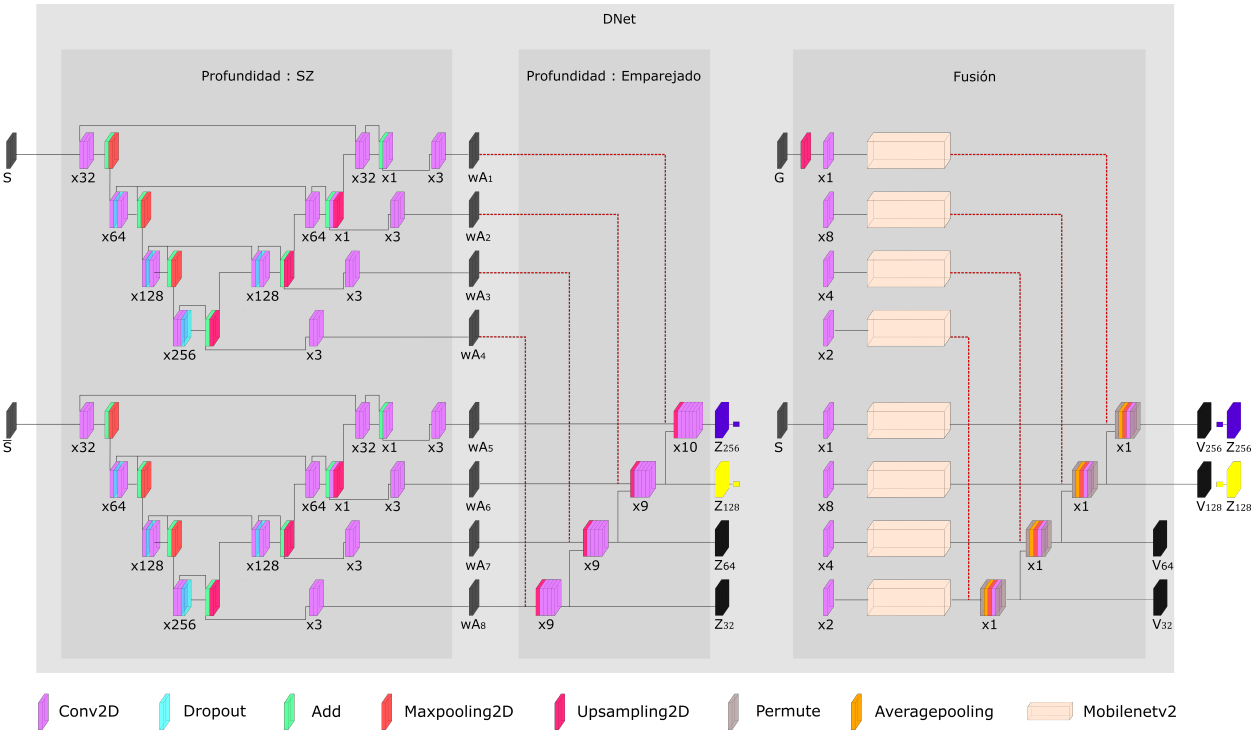

Figure 5.1. Diagrama de flujo de la red DNet que parte de 3 entradas que tienen resolución espacial de 256×256, pasa por los 4 bloques principales de la red y se obtiene la salida de una mapa de profundidad **Z** y un imagen VNIR **V**.

ling, hasta llegar a un determinado tamaño, a partir del cual la red comienza a presentar salidas a distintas resoluciones dando como resultado salidas como A1, A2, A3 y A4 de dimensiones de resolución espacial de $256 \times 256 \times 1$, $128 \times 128 \times 1$, $64 \times 64 \times 1$ y $32 \times 32 \times 1$ respectivamente, dichas salidas a su vez son comparadas con los bloques de convolución respectivos del brazo derecho e izquierdo de la red unet, es decir con los bloques cuyas salidas tienen la misma resolución. Específicamente los bloques S-Z se componen por 25 convoluciones, 3 bloques de submuestreo, 4 bloques de dropout, 3 bloques de sobremuestreo, donde cada implementación de estos bloques tienen definido un número de parámetros entrenables. Luego de pasar por los bloque S-Z, las salidas de los bloques S-Z-1 y S-Z-2 ingresan al bloque de emparejado 5.4 compuesto por una capa que

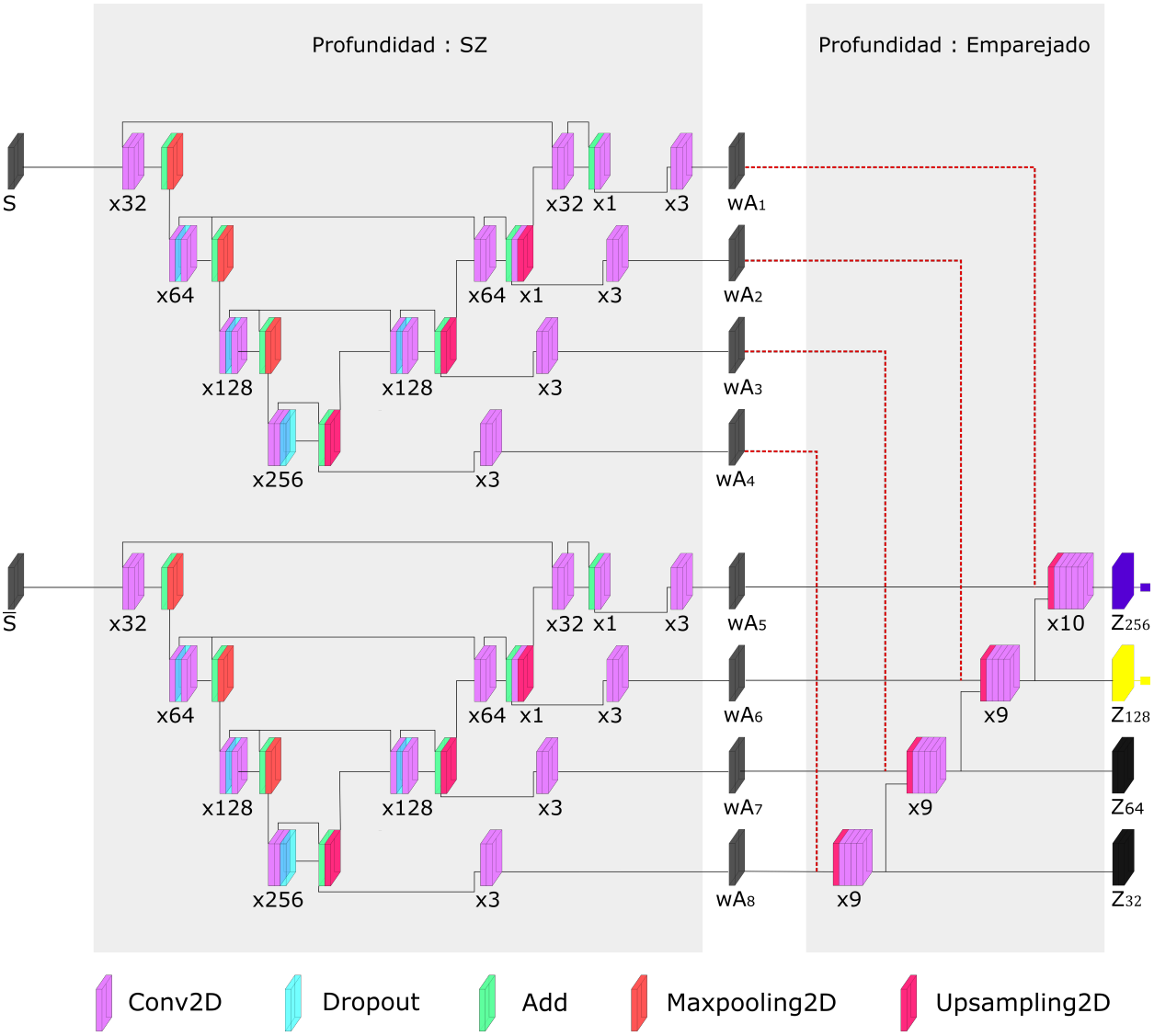

Figure 5.2. En esta imagen se presenta la estructura del bloque completo de profundidad que cuenta con dos bloques internos Profundidad: SZ y Profundidad: Emparejado

básicamente implemente 18 convoluciones, 4 concatenaciones y 3 sobre muestreos. En este bloque, se generan las salidas de los mapas de profundidad. Durante este proceso de realiza una serie de concatenaciones, convoluciones y muestreos. En este proceso, se inicia con una concatenación de las salidas de los bloques S-Z de menor resolución a los cuales se les realiza una serie de cuatro

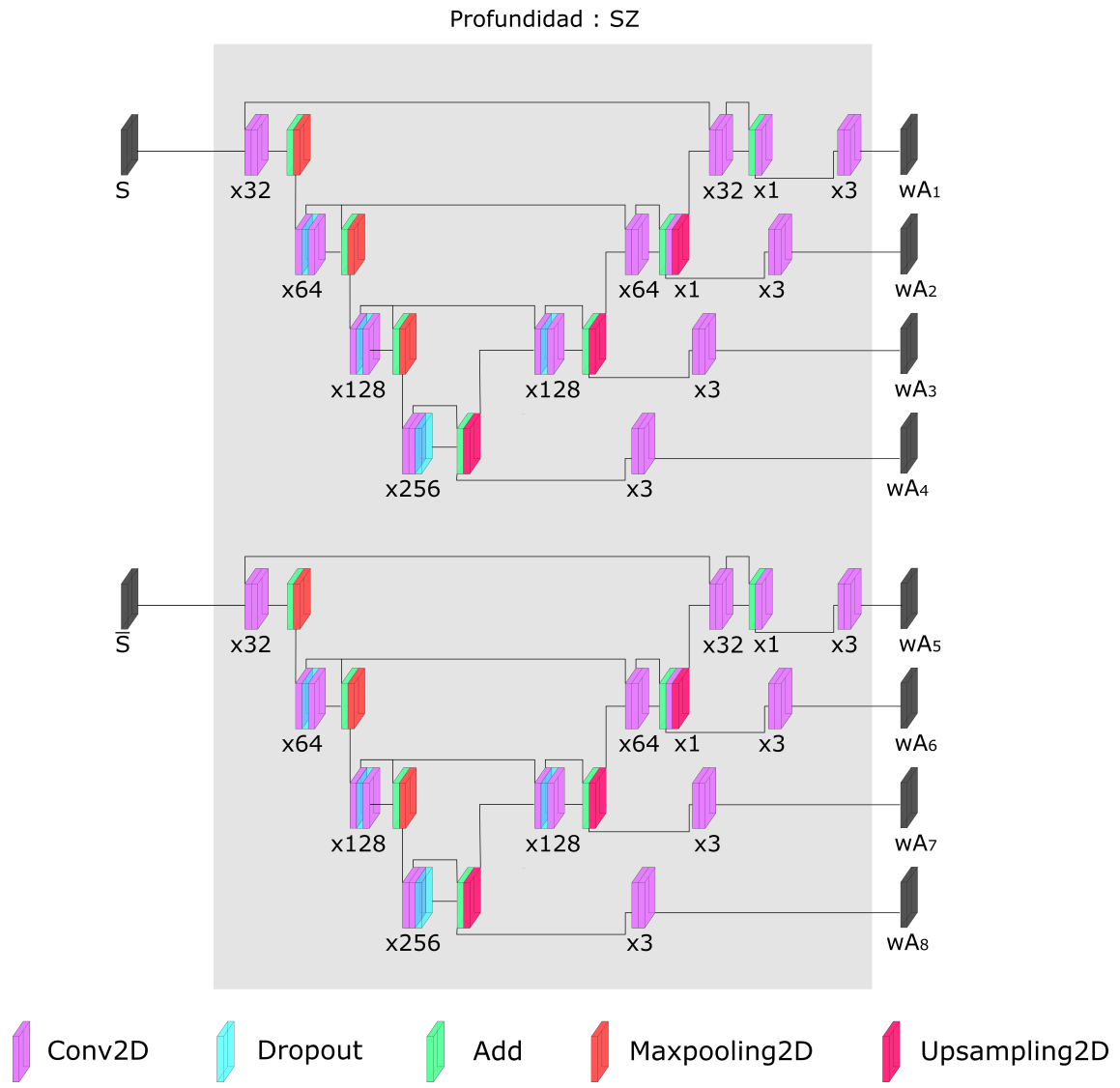

Figure 5.3. En esta ilustración se presenta el bloque de profundidad: SZ que sigue la estructura de una red Unet .

convoluciones, y luego este resultado pasa por un bloque de sobre muestreo de 2×2 y se concatena con las salidas S-Z de resolución espacial de 64×64.Este proceso de subida continua hasta realizar las convoluciones para las salidas de los bloques S-Z de resolución espacial de $256x256$.

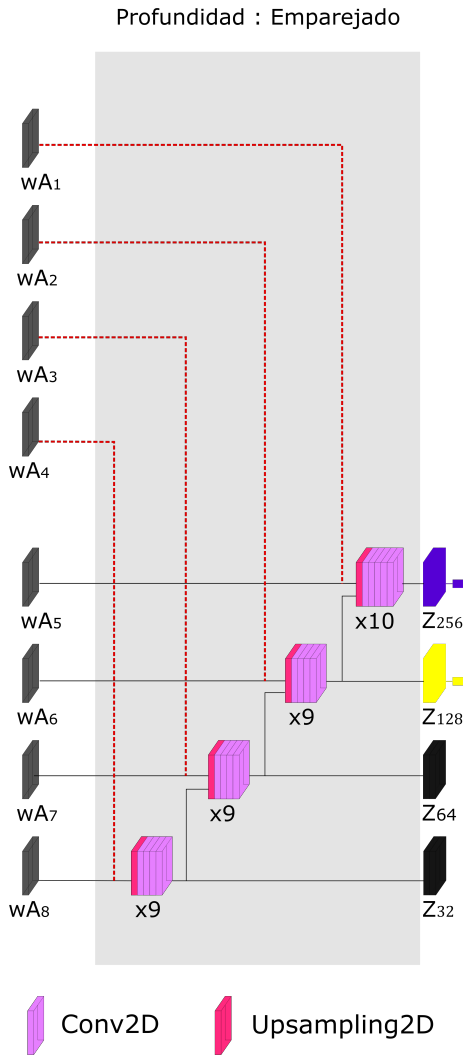

Figure 5.4. En este imagen se presenta el bloque de profundidad de emparejado que es el bloque en donde las imágenes son concatenadas para generar los mapas de profundidad.

5.1.2. Bloque Fusión. El bloque de fusión se encarga de realizar la fusión de una imagen capturada por el sensor derecho **S** y la imagen capturada por el sensor infrarrojo **G**, donde la principal intensión es obtener una imagen RGB-NIR **V**. Para ello este bloque fusiona imágenes de resoluciones espaciales de 256×256 y 64×64 respectivamente. Este bloque se compone de

modulo de fusión y el modulo de atención de canal, donde el modulo de fusión, es una arquitectura basada en la mobileNetv2 [Sandler et al., 2018] con los pesos cargados de la base de datos ImageNet como punto inicial. Particularmente en el bloque de fusión se realizan 3 submuestreos y 4 convoluciones. Este modulo tiene 4 salidas de imágenes RGB-Nir de resoluciones espaciales de 256×256, 128×128, 64×64 y 32×32. Luego estas imágenes ingresan a un bloque de atención con estructura idéntica al propuesto por el trabajo de [Zhang et al., 2020] permiten seleccionar las características de un conjunto, y mitigar la redundancia. Específicamente, este bloque esta compuesto por una capa de permutación a la entrada, dos capas convoluciones dos bloque adicionales de max pooling y average pooling y nuevamente dos bloque convoluciones de upsampling y adicionalmente una capa de permutación que permite tener como salida una imagen del mismo tamaño en la salida y en la entrada.

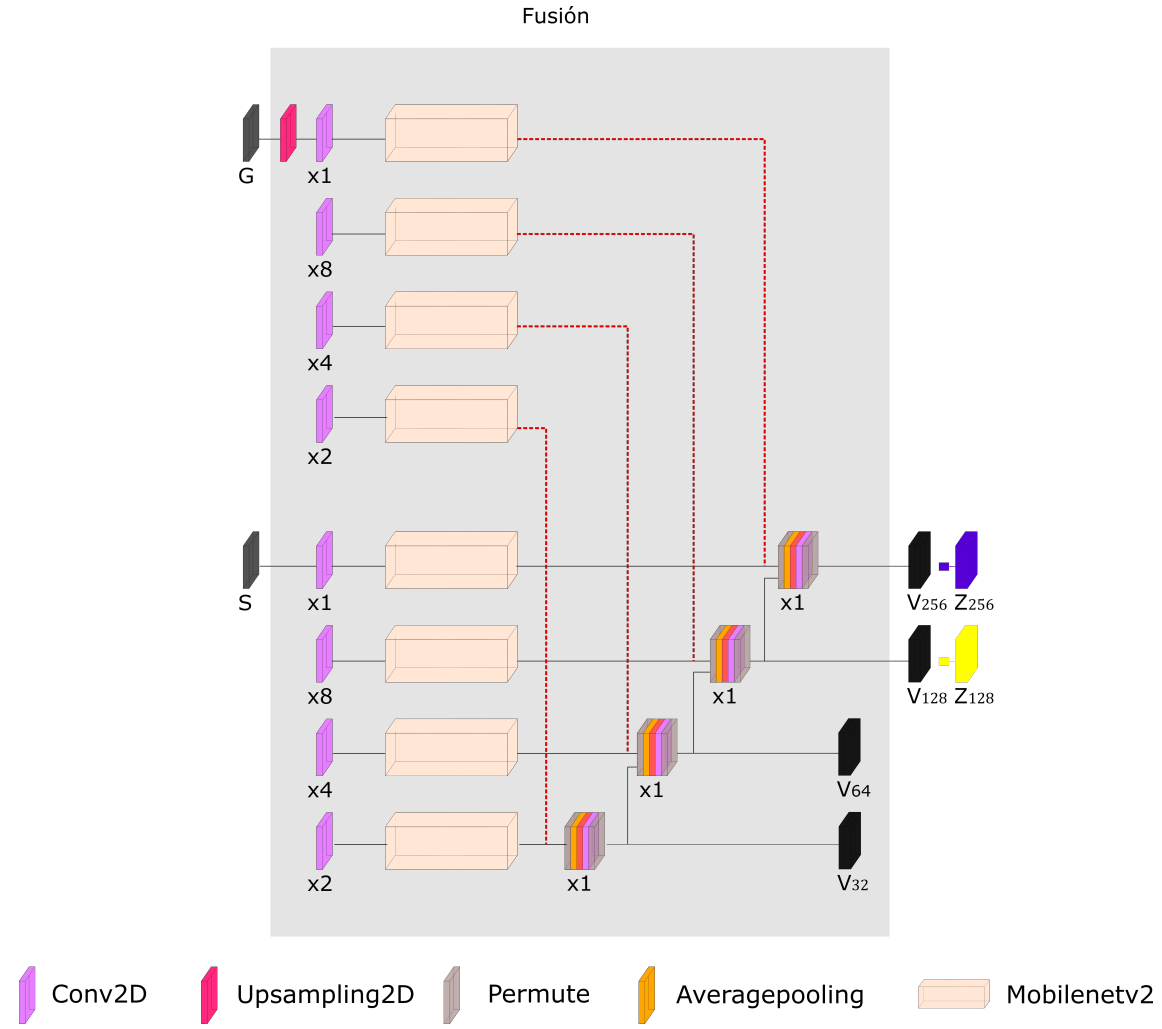

Figure 5.5. En esta imagen se presenta el bloque de fusión que esta compuesto principalmente por un bloque principal de fusión que consta de la red mobilenetv2 y un canal de atención presentado en el estado del arte. Fuente: Propia.

6. Resultados y simulaciones

6.1. conjunto de datos

El conjunto de datos elegido para el entrenamiento de la red propuesta D-Net, es el conjunto de datos de estero visión propuesto por Middleburry College, que fue adquirido utilizando una arquitectura de estéreo visión portátil que cuenta con dos cámaras Canon DSLR (EOS 450D con lente de 18–55 mm) en modo de resolución media (6 MP) y dos cámaras de apuntar y disparar, que fueron montadas sobre un carril óptico horizontal con línea base variable de 140 mm a 400 mm que combinado con la técnica descrita en [Scharstein et al., 2014]. Las capturas de e conjunto de datos fueron tomadas bajo ambiente de laboratorio controlado, lo que dio paso a la generación de capturas con alta precisión de disparidad, con datos como los contenidos en la tabla 6.1.

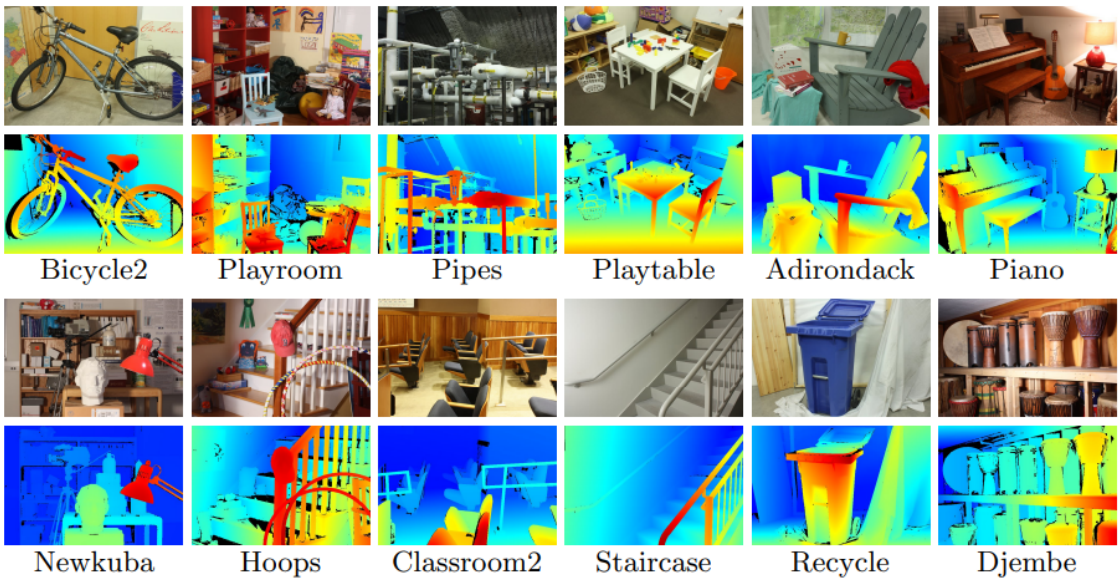

Figure 6.1. Data conjunto Middlebury. Fuente: D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth.

Nombre	**Característica de la escena**
SCENE-{perfect,imperfect}	Calibración perfecta e imperfecta por escena
ambient	Directorio de todas las vistas de entrada bajo iluminación ambiental
L{1,2,...}	Diferentes condiciones de iluminación
im0e{0,1,2,...}.png	Vista izquierda bajo diferentes exposiciones
im1e{0,1,2,...}.png	Vista correcta bajo diferentes exposiciones
calib.txt	Información de calibración
im{0,1}.png	Vista izquierda y derecha por defecto
im1E.png	Vista izquierda y derecha por defecto bajo diferentes exposiciones
im1L.png	Vista derecha predeterminada con diferente iluminación
disp{0,1}.pfm	GT de disparidades izquierda y derecha
disp{0,1}-n.png	GT izquierdo y derecho número de muestras (* solo perfecto)
disp{0,1}-sd.pfm	GT con desviaciones estándar izquierda y derecha (* solo perfectas)

Tabla 6.1. Tabla de descripción de los datos brindados por el data conjunto de esterero visión de Middlebury

Específicamente para la evaluación de la red DNet, se utilizaron los archivos que corresponden a la imagen izquierda y derecha capturado por el sensor, al archivo de texto de la arquitectura con

los parámetros necesarios para calcular el mapa de disparidad con referencia en la escena derecha o izquierda. Dichos datos fueron preprocesados y ajustados con el proceso descrito en la siguiente sección.

6.2. Preprocesamiento

Se utilizó el grupo de datos de Middlebury College para construir un conjunto de **36800** archivos de entrenamiento, cada una de las imágenes RGB fue normalizadas y llevadas al tamaño 256×256. A partir de los mapas de profundidad, se generaron los mapas de disparidad con una capa personalizada. Con los mapas de disparidad y las imágenes RGB como la imagen adquirida por la cámara derecha, se generaron las imágenes izquierdas, sin embargo, debido a que el bloque generador de imágenes izquierdas tiene una alta complejidad computacional y depende de la distancia entre cámaras, que es un dato aprendible, se optó por dejar el bloque de generación de imágenes izquierdas como una etapa de preprocesamiento, fijando 4 posibles distancias entre cámaras. Para cada una de las distancias máximas de las imágenes del grupo de imágenes de entrenamiento, la profundidad de los mapas fueron escaladas a 1, 5, 10 y 15 metros. En general se generó un grupo de datos de entrenamiento que contiene 2300 imágenes derechas, 9600 mapas de profundidad, que se dividen en 4 grupos que dependen de la distancia máxima con 2300 mapas de profundidad por grupo , y 36800 imágenes izquierdas. Además es importante destacar que debido a que no se encontraron conjuntos de datos, que contuvieran al mismo tipo información de profundidad e infrarrojo, se optó por generar sintéticamente estas últimas, promediando las 3 bandas de intensidad de las imágenes RGB derechas ya mencionadas, posteriormente se les realizó un escalado espacial.

6.2.1. Escala de mapas de profundidad, recorte y cambio de resolución espacial. Se parte de las imágenes derechas del conjunto de datos (S), el mapa de disparidades (D), la distancia entre las cámaras con las que se capturaron las imágenes del conjunto de datos (β), la distancia focal (f) el parámetro *offset* en las disparidades ($doffs$), y de la ecuación

$$Z = \beta * 10^{-3} * f/(doffs + D) \tag{6.1}$$

se calcula el mapa de profundidad Z en una resolución espacial de 1920×1080, posteriormente tanto Z como S se escalan a una resolución espacial de 512×512. Posteriormente se pasa a un ciclo que normaliza el mapa de profundidad con respecto al máximo y lo multiplica por cada una de las 4 distancias máximas (1m, 5m, 10m, 15m), para cada una de estas distancias se generan 100 recortes de resolución de 256×256 a partir de puntos aleatorios en S y Z. Además se genera la imagen infrarroja G simulada promediando las 3 bandas de color de S, G y se submuestrean a 64×64 S para generar las imágenes V.

6.2.2. Algoritmo para generar imágenes izquierdas.. El algoritmo 1 presenta el proceso, que genera imágenes izquierdas $\overline{S}$ a partir de la definición de disparidad (2.2). Este algoritmo fue programado como una capa lambda. La lógica que rige este algoritmo es recorrer el mapa de disparidad por todas sus filas y columnas, y desplazar el píxel correspondiente que se esta analizando hacia la derecha (debido que nuestra imagen base es S) una cantidad de píxeles igual al valor del mapa de disparidad en el punto, se muestra en el algoritmo de la linea 8 a la linea 13, en estas lineas se describe una mascara que de forma secuencial barre de izquierda a derecha una

fila, para posteriormente saltar a la siguiente , en la linea 14 se realiza una multiplicación punto a punto denotada por el operador $\odot$.

Algorithm 1 Estimación de imágenes izquierdas a partir de imágenes derechas y mapas de disparidad

Input: S ,D
// S es la imagen derecha y D el mapa de disparidad
Output: $\overline{S}$

1: $N \leftarrow length(S)$ // Se toma el tamaño de la imagen
2: $MASK \leftarrow [Zeros(N,N,3)]$ // Se genera una mascara binaria
3: $MASK[0,0,:] \leftarrow 1$ // La esquina superior izquierda se pone en 1
4: $CTE \leftarrow max(D)$ // Se calcula el máximo que se desplaza un píxel
5: $MASK \leftarrow [MASK, Zeros(N,CTE,3)]$// Se expande la mascara
6: $L_{im} \leftarrow [Zeros(N,N+CTE,3)]$ // Se inicia la imagen en 0
7: $i \leftarrow 0$
8: **for** $i \leq N-1$ **do**
9: $j \leftarrow 0$
10: $MASK1 \leftarrow roll(MASK,i,0)$ // La función roll rota i posiciones la matriz en direccion de ascenso de las columnas
11: **for** $j \leq N-1$ **do**
12: $Rol \leftarrow roll(S,D[i,j],1)$// La función roll rota la matriz la disparidad en el píxel
13: $MASK2 \leftarrow roll(MASK1, j+D[i,j],1)$
14: $L_{im} \leftarrow L_{im} + MASK2 \odot Rol$ // Se suma la imagen con la multiplicación punto a punto de la mascara binaria rotada con la imagen rotada, para afectar solo 1 píxel
15: **end for**
16: **end for**
17: $\overline{S} \leftarrow Lim[0:N-1, 0:N-1, :]$ // Se recorta la imagen de salida a las dimensiones necesarias

6.3. Simulaciones

Los bloques propuestos en la red DNet fueron evaluados para entrenamiento utilizando datos preprocesados como se mencionó en la sección 6.2. Específicamente, se realizaron distintas evaluaciones teniendo en cuenta variaciones de profundidad $Z = 1,5,10,15[m]$ y distancia entre cámaras

$\beta = 0{,}01, 0{,}1, 0{,}5, 1[m]$. En cada uno de estos entrenamientos se implementaron un conjunto de 2300 imágenes de las cuales 1840 fueron usados en entrenamiento y 460 en validación. Para la evaluación de las métricas obtenidas en la prueba se implemento un conjunto de imágenes que no fueron implementadas en ningún momento del entrenamiento.

6.4. Descripción de parámetros de entrenamiento

Para el entrenamiento de la red DNet propuesta se toma el conjunto de datos preprocesados descritos en la sección 6.2. Y se realiza el entrenamiento de 16 redes DNet en forma paralela utilizando distintos conjuntos de datos, teniendo en cuenta que cada red es la combinación de los los valores de distancia entre cámaras $\beta = \{0{,}001, 0{,}1, 0{,}5, 1\}[m]$ y los rangos de profundidad $\mathbf{Z} = \{1, 5, 10, 15\}[m]$. Cada red DNet fue evaluada, utilizando un lote de entrenamiento $BatchSize = 8$, una taza de aprendizaje de $L_t = 10^{-3}$, una función de costo error cuadrático medio (ECM)y fue entrenada un aproximado de 5.000 épocas cada una. Estás redes fueron sometidas a entrenamiento un aproximado de 2 semanas consecutivas, utilizando primero, el entorno colaborativo de google *COLAB*, dos GPUs distintas con las siguientes especificaciones: la primera, con una tarjeta gráfica Quadro RTX 8000, con Procesador Intel(R) Xeon(R) W-3223, una CPU @ 3.50GHz x 16 y 93GB de memoria RAM; y la segunda con una tarjeta gráfica GeForce RTX 3090 X 2 , un procesador, Intel(R) Xeon(R) W-3223, y una CPU @ 3.50GHz.

6.5. Resultados de la simulación

Para la evaluación en simulación del rendimiento de la DNet, se presentan las resultados de la predicción de los mapas de profundidad e imagen infrarroja, con respecto a las imágenes de referencia tanto para imágenes de entrenamiento como para imágenes de validación,teniendo en cuenta las

métricas de PSNR y SSIM.

	Imágenes de Entrenamiento							
	Profundidad 1 m				**Profundidad 5 m**			
	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$
PSNR Profundidad	31.08	**32.37**	28.64	27.01	31.29	31.30	**32.25**	31.12
SSIM Profundidad	0.9252	**0.9387**	0.8861	0.8653	0.7355	0.7407	**0.7798**	0.7453
PSNR IRGB	32.93	**32.96**	**32.96**	32.93	32.95	32.93	**32.96**	32.94
SSIM IRBG	0.97202	**0.97213**	0.97211	0.97211	0.97184	0.97181	**0.97188**	0.97185
	Profundidad 10 m				**Profundidad 15 m**			
	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$
PSNR Profundidad	29.44	**31.32**	29.80	29.76	**29.75**	28.94	29.65	29.65
SSIM Profundidad	**0.6077**	0.5955	0.6036	0.5905	**0.4581**	0.4479	0.4375	0.4363
PSNR IRGB	**32.95**	32.93	32.94	**32.95**	32.95	32.95	**32.96**	32.93
SSIM IRBG	0.97203	0.97204	**0.97214**	0.97204	0.97193	0.97193	**0.97196**	**0.97196**

Tabla 6.2. Métricas promedio para imágenes de entrenamiento.

Específicamente, los datos mostrados en la Tabla.6.2 muestran los resultados para las imágenes de entrenamiento. Particularmente se ve que para el nivel de profundidad de 1 metro, los mejores resultados en la reconstrucción de profundidad y de color se obtuvieron para la distancia entre cámaras de 0.1, que obtuvo entre $0{,}57 - 5{,}36[dB]$ mas que las otras distancias. Para el caso de la profundidad de 5 metros, se tuvo que la mejor distancia entre cámaras fue la distancia de 0.5 metros ganando en ambas métricas PSNR y SSMI para profundidad y color donde específicamente para la profundidad se tiene que gano por entre $0{,}95 - 1{,}13[dB]$. Ahora, para el caso de una profundidad de 10 metros, se obtuvo que para la métrica de PSNR de profundidad la distancia entre cámaras que tuvo el mejor desempeño fue la de 1 metro, sin embargo, para la métrica de SSIM de profundidad la que tuvo mejor desempeño fue la menor distancia es decir la de 0.01 metro, además, en cuanto a las métricas de evaluación de color, la mejor reconstrucción de color la dio la distancia entre cámaras de un metro. Y para la profundidad de 15 metros se tiene que esta tuvo un mejor desempeño para ambas métricas en profundidad para la distancia entre cámaras de 0.01 metro, y para las métricas en color para una distancia entre camas de 0.5 metros. Finalmente, se logra evidenciar que la mejor reconstrucción teniendo en cuenta distancia entre cámaras y profundidad, se obtuvo a una distancia de 0.1 metro entre cámaras a una profundidad de 1 metro, sin embargo, cabe resaltar que si se revisa con detenimiento todos los resultados de la tabla, se puede notar que la distancia que tuvo resultados mas estables para todas las distancias entre cámaras fue la profundidad de 5 metros. Algunos ejemplos que evidencian lo anterior pueden ser observados en las Figuras 6.2, 6.3,6.4, 6.5 La evaluación de las imágenes de prueba se observa en la Tabla 6.3, estos resultados muestran que para la profundidad de 1m, la distancia entre cámaras que obtuvo el mejor rendimiento en la

	Imágenes de Prueba							
	Profundidad 1 m				**Profundidad 5 m**			
	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$
PSNR Profundidad	15.36	14.97	14.91	**15.72**	12.70	14.50	14.89	**15.39**
SSIM Profundidad	**0.7161**	0.6858	0.6486	0.6517	0.2532	0.3796	**0.3849**	0.3523
PSNR IRGB	32.19	32.18	**32.20**	**32.20**	32.18	**32.19**	32.18	32.16
SSIM IRBG	0.98158	0.98158	**0.98160**	0.98159	0.98156	0.98156	**0.98157**	0.98154
	Profundidad 10 m				**Profundidad 15 m**			
	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$	$\beta = 0{,}01$	$\beta = 0{,}1$	$\beta = 0{,}5$	$\beta = 1$
PSNR Profundidad	12.40	14.47	15.09	**15.29**	12.52	14.25	12.47	**15.35**
SSIM Profundidad	0.1377	0.2572	**0.2666**	0.2505	0.0843	0.1270	0.0899	**0.1917**
PSNR IRGB	32.18	32.18	32.17	**32.20**	**32.21**	**32.21**	32.20	32.18
SSIM IRBG	0.98158	0.98159	0.98157	**0.98160**	0.98160	0.98160	**0.98161**	0.98156

Tabla 6.3. Comparación del rendimiento promedio para imágenes de validación variando el parámetro de distancia de cámaras, en términos de las métricas PSNR y SSIM.

métrica de PSNR es la distancia de 1m, superando a las demás distancias entre $0{,}36 - 0{,}81[dB]$. Ahora, para el nivel de profundidad de 5 metros, se tiene que la distancia entre cámaras que tiene el mejor resultado para el PSNR de profundidad, es la distancia de 1 metro, mejorando entre $0{,}5 - 2{,}69[dB]$. De igual manera, para las profundidades restante de 10 metros y 15 metros, se tiene que la mejor distancia entre cámaras es la de 1 metro. Lo anterior permite concluir, que distancia óptima para las imágenes de validación entre cámaras para todos los niveles de profundidad es de 1 metro. Algunos ejemplos que evidencian lo anterior pueden ser observados en las Figuras 6.6, 6.7,6.8, 6.9.

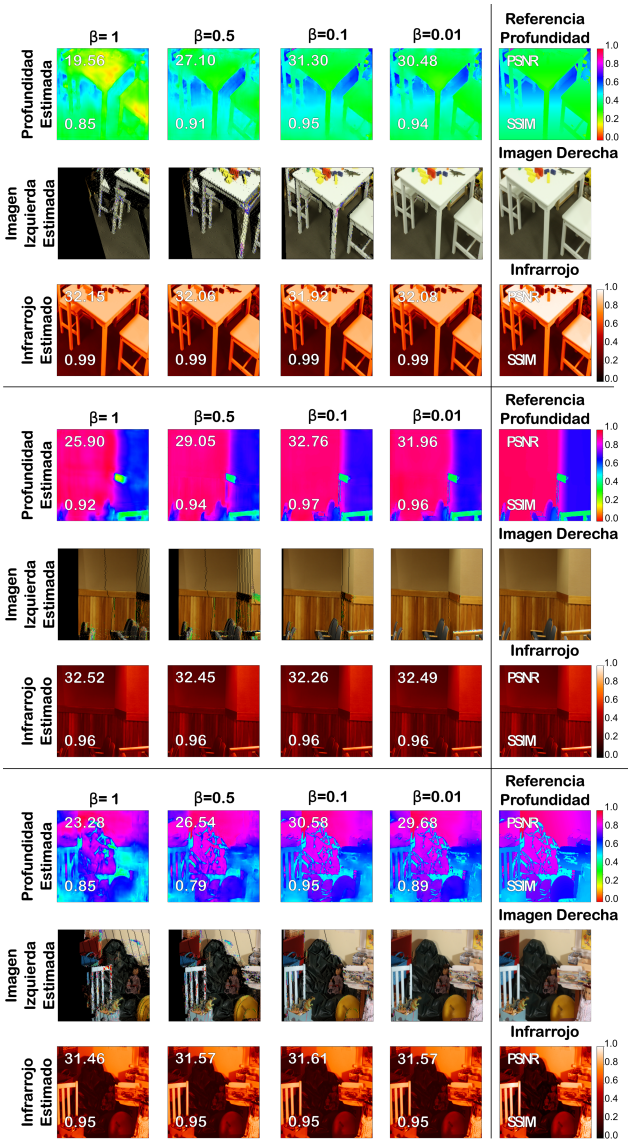

Figure 6.2. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de entrenamiento para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 1m

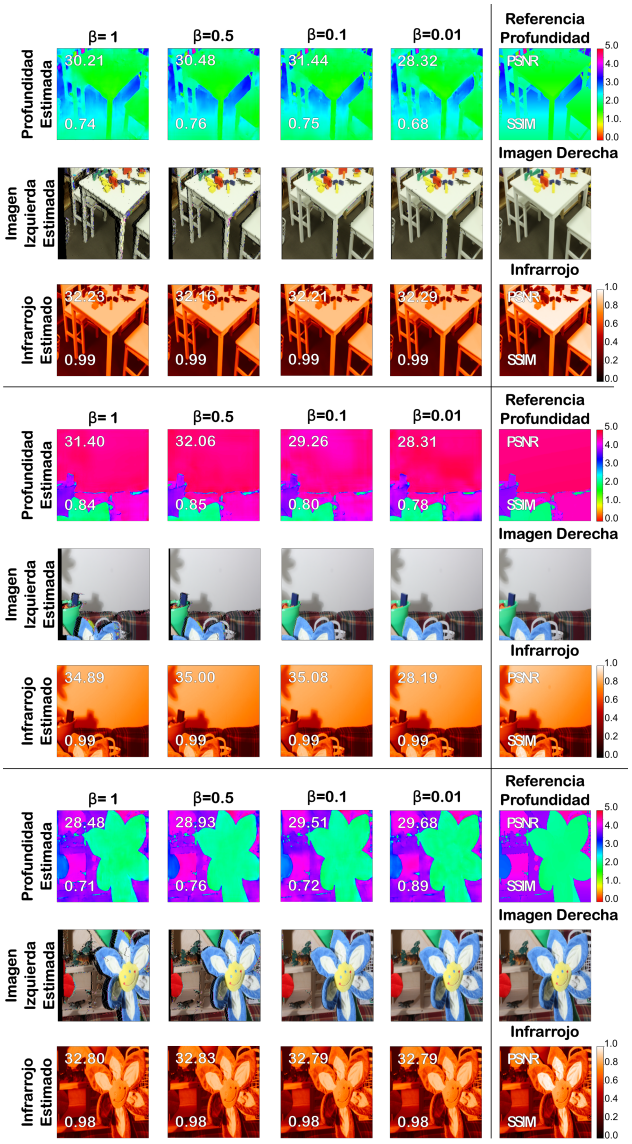

Figure 6.3. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de entrenamiento para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 5m

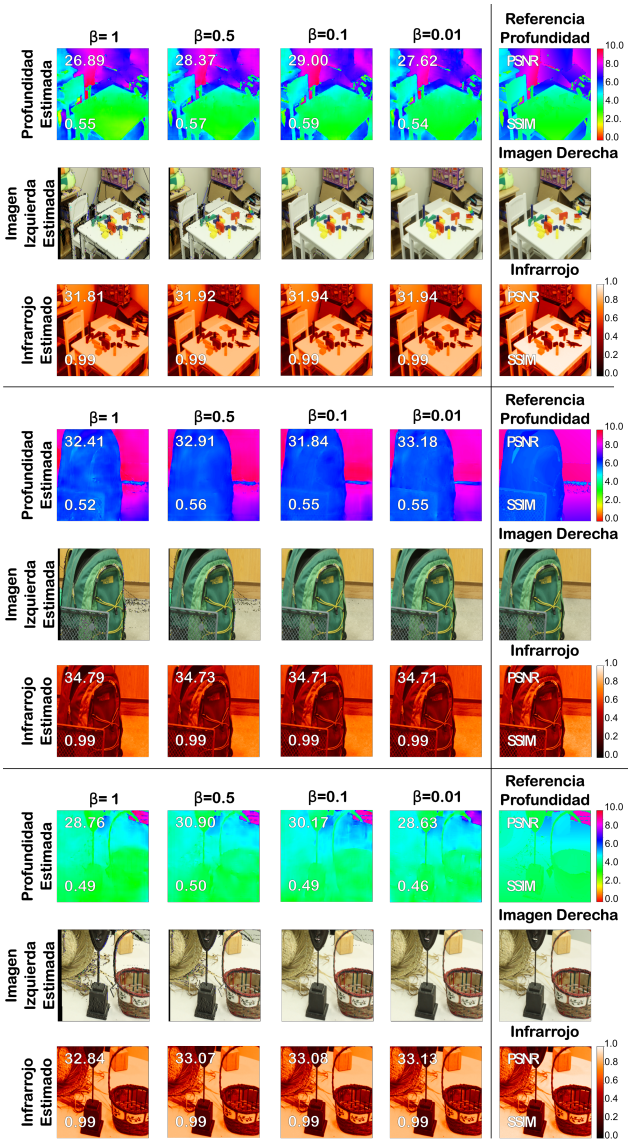

Figure 6.4. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de entrenamiento para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 10m

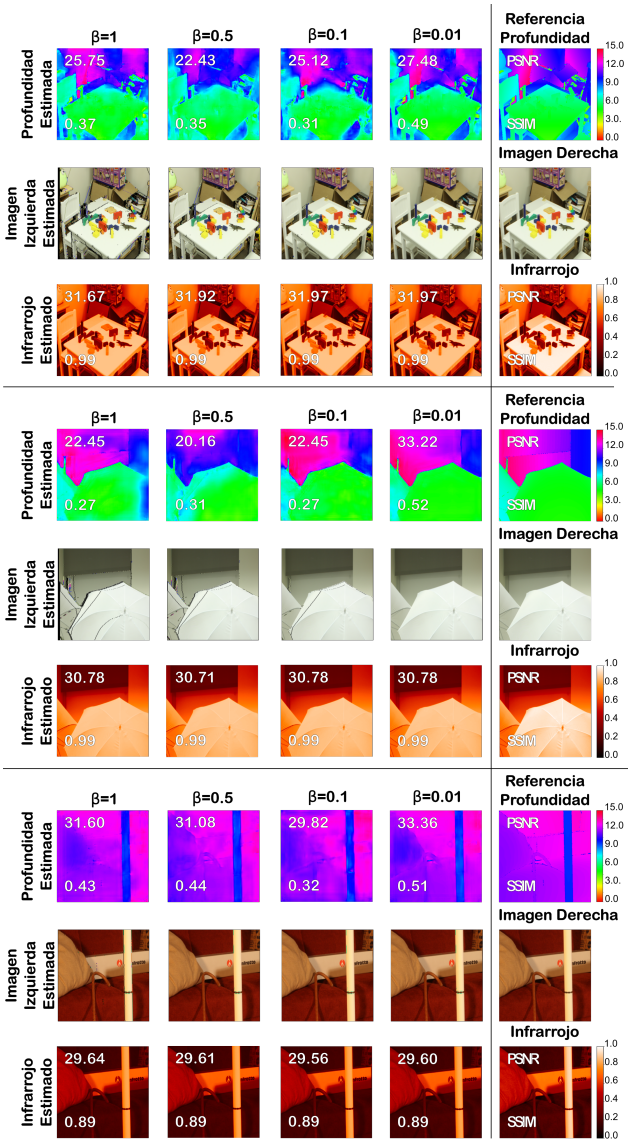

Figure 6.5. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de entrenamiento para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 15m

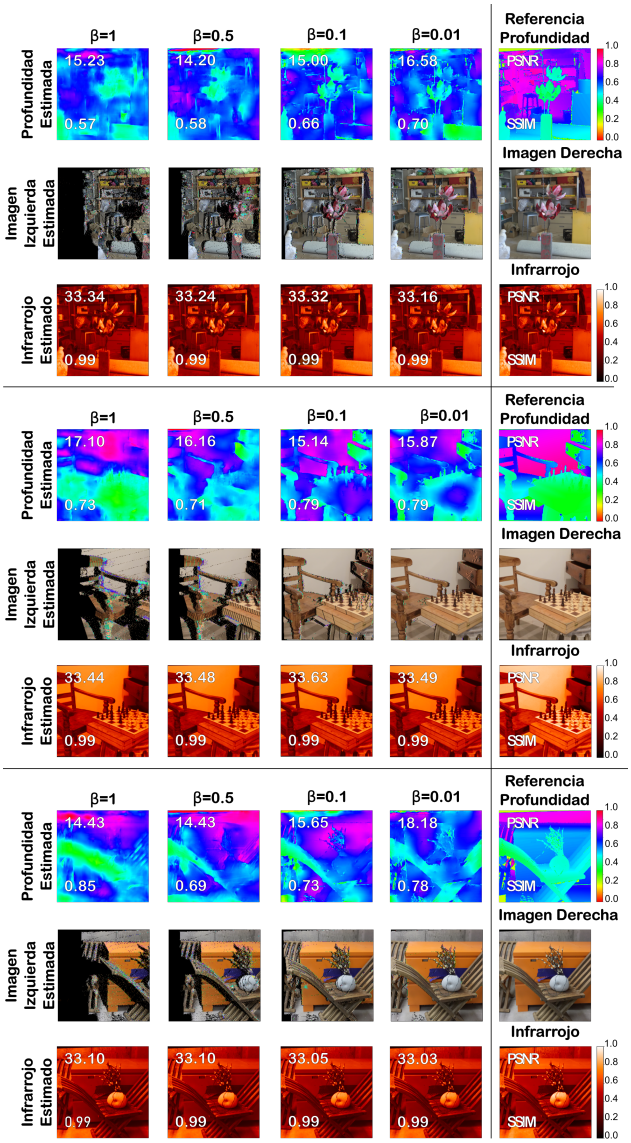

Figure 6.6. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de validación para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 1m

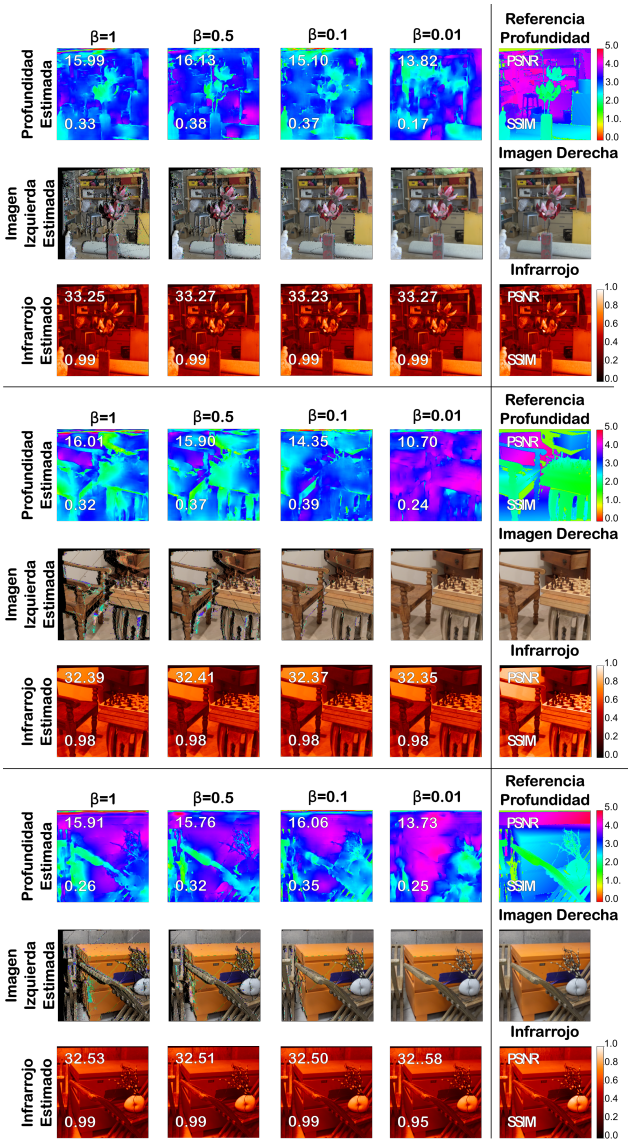

Figure 6.7. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de validación para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 5m

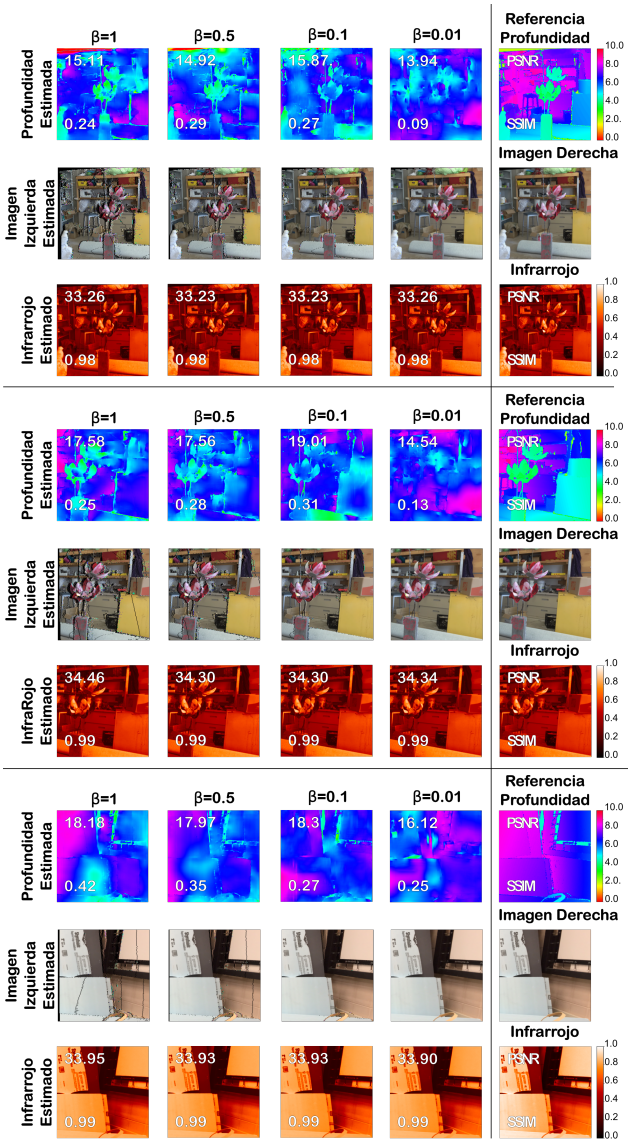

Figure 6.8. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de validación para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 10m

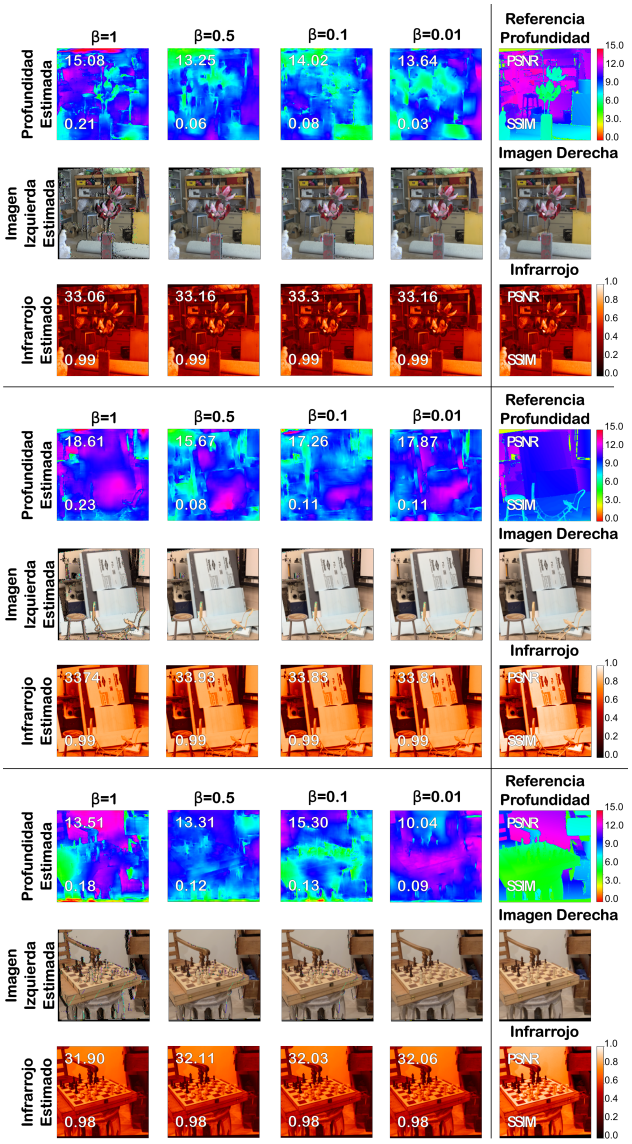

Figure 6.9. Comparativa visual entre las predicciones de DNet y las referencias en imágenes de validación para las métricas PSNR y SSIM tanto para profundidad como para las imágenes IRGB, con distancias entre cámaras de 1cm, 10 cm, 50cm y 1 metro y una profundidad máxima de 15m

7. Implementación del sistema multisensor

En sistema multisensor fue evaluado en un ambiente no controlado, es decir que sus capturas fueron realizadas utilizando un espacio abierto con luz natural. Espacio en el cual se tomaron 14 escenas distintas realizando variación de distancias entre cámaras, distancias entre objetos, además de la variación natural de la iluminación.

7.1. Sistema Óptico

El montaje multisensor implementando consta de 2 sensores electrónicos de marca Logitech con referencia c922 de resolución 720p que son los encargados de capturar la información RGB de las imágenes izquierdas. Además se cuenta con una cámara de seguridad que cuenta con un sensor infrarrojo y RGB para la captura del sensor de la derecha, la marca de este elemento TP-LINK de referencia tapo c200 de resolución 1080p.

Figure 7.1. en esta imagen se presenta el montaje utilizado para implementación y captura de las imágenes de evaluación de la red, este montaje cuenta con dos cámaras de RGB y una cámara de seguridad de RGB/ infrarrojo.

El montaje se compone por un riel de metal, con referencias para las distancias entre cámaras de 1cm, 10cm, 50cm y 100cm. En la Figura. 7.1 se observa en detalle el montaje realizado siendo la cámara de la izquierda el sensor de la marca TP-LINK y las cámaras de la derecha los sensores de la referencia c922 de logitech.

7.1.1. Descripción. Con el fin de capturar escenas con el sistema óptico multisensor, se toma como primer paso la configuración del sistema presentado en la Figura , en este sistema se observan tres cámaras que se encuentran sobre un ángulo que actúa como riel, por el cual se van a desplazar las cámaras. Teniendo en cuenta que las cámaras que realizan el movimiento son las cámaras RGB de la misma referencia y la cámara IR permanece estática se presentan a continuación la configuración de las escenas.

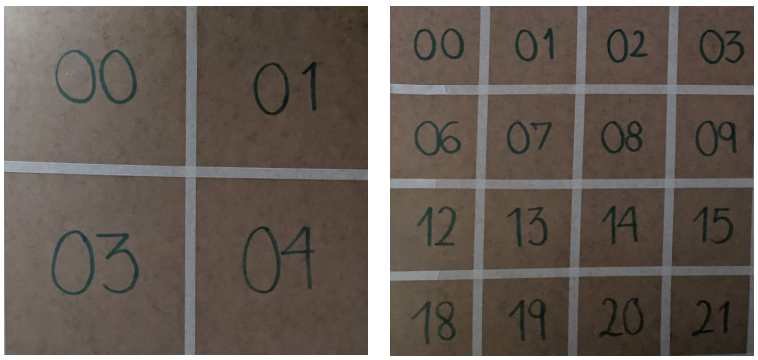

Figure 7.2. Cuadros objetivos para calibrar el sistema, el cuadro de la izquierda presenta un tamaño de recuadros de 10cm$\times$ 10cm, el cuadro de la derecha presenta un tamaño de recuadros de 20cm$\times$ 20cm

Para la captura de las escenas, se fabrica el objetivo de la Figura. 7.1 , que es un rejilla numerada que se utiliza para asegurar la calibración del sistema. Luego de la calibración, se procede a tomar 14 escenas en donde aparecen diversos objetos en distintas posiciones (x, y, z), teniendo en cuenta que para cada escena se realizan dos capturas por cada sensor en un plano de la misma escena en donde se mide la posición a la que se encuentra cada objeto con respecto a la cámara que además de las dimensiones del mismo. En total se capturan las 14 escenas teniendo en cuenta para las primeras 5 capturas se tiene una distancia máxima de 1 m en profundidad, y una distancia máxima

de 15 m aproximadamente para las capturas restantes.

7.2. Resultados Predicción

Para la evaluación del rendimiento de la red DNet, primero se capturan imágenes con el sistema multisensor propuesto, bajo condiciones de iluminación natural, en un ambiente no controlado. Luego se realiza un pre-procesamiento de los datos, que consiste, en un ajuste de color de las imágenes izquierda y derecha, donde se ajustó el histograma para que ambas imágenes siguieran un comportamiento similar. Luego se realizaron recortes a la imagen capturada de la escena, ajustando el tamaño a las dimensiones necesarias para ingresar a la red DNet. Después dichas imágenes son evaluadas por la red DNet,y comparadas teniendo en cuenta que las profundidades de referencia que fueron adquiridas mediante mediciones del objeto respecto a la cámara del infrarrojo utilizando un flexómetro. Por tanto, se compara la imagen de resultado de la predicción con la imagen de la escena capturada por la cámara de infrarrojo de referencia y finalmente, se encuentra que para las capturas tomadas bajo una ambiente con iluminación natural y en un ambiente no controlado, la red DNet, tiene una predicción aceptable como puede observarse en las Figura 7.3 y Figura 7.4.

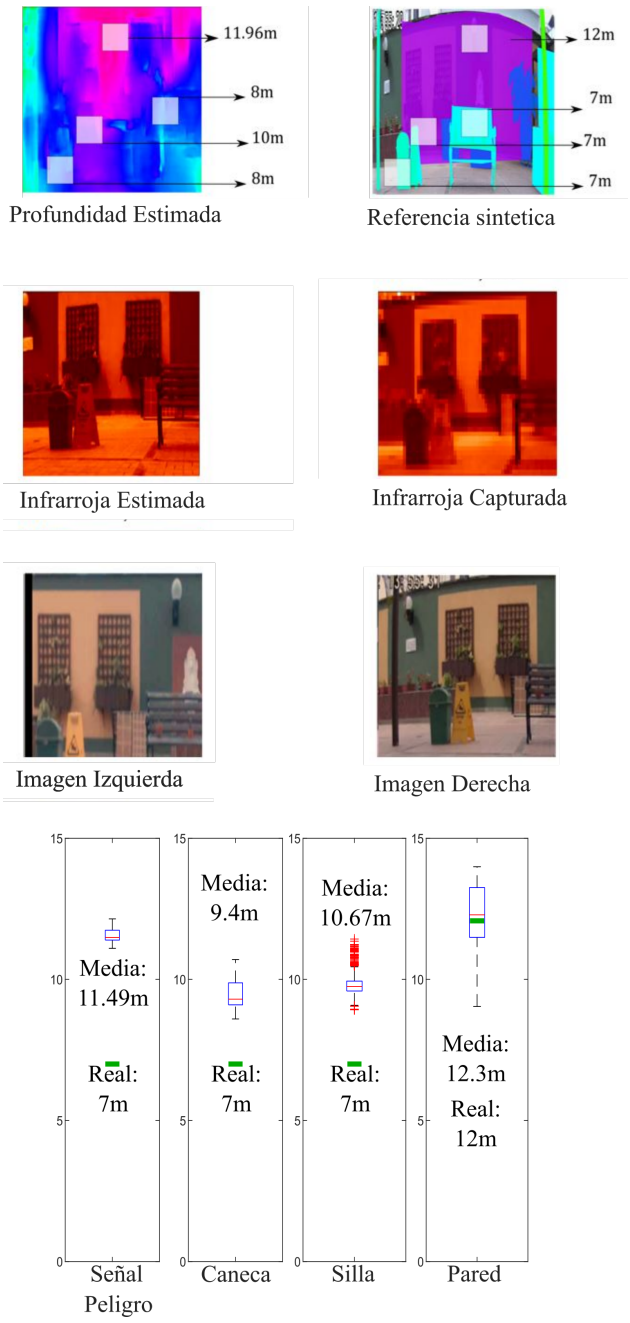

Figure 7.3. Resultado de una predicción hecha por la Red DNet para una escena capturada en un ambiente no controlado con sus respectivas imágenes de referencia

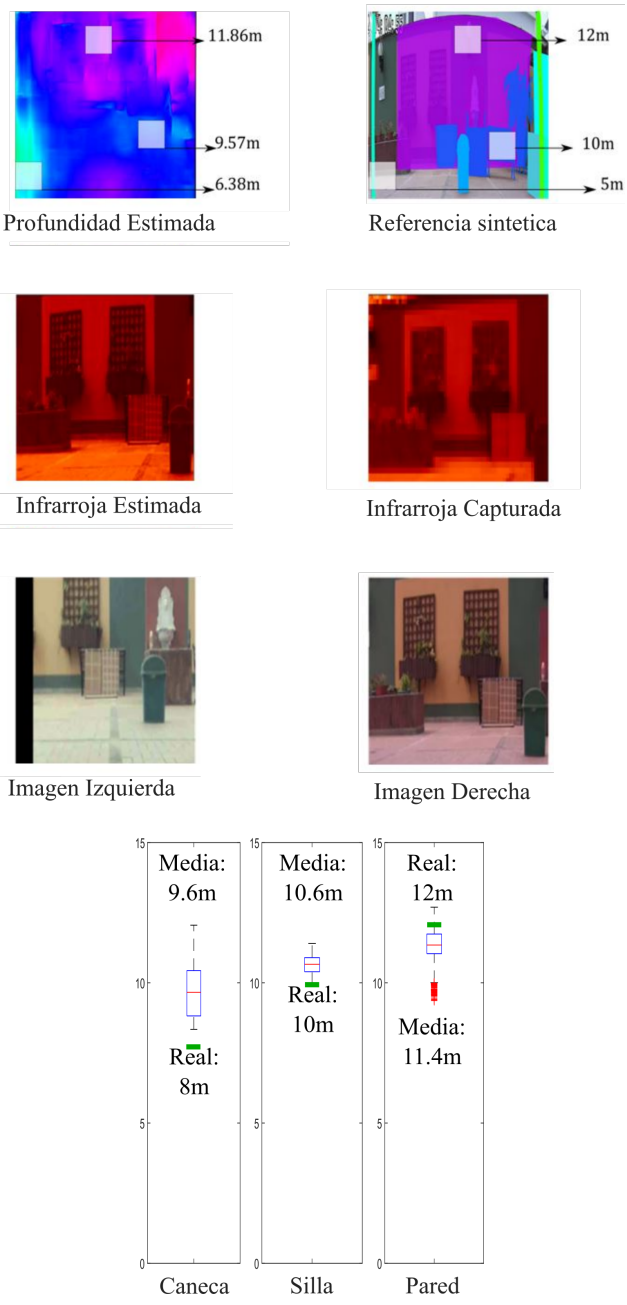

Figure 7.4. Resultado de una predicción hecha por la Red DNet para una escena capturada en un ambiente no controlado con sus respectivas imágenes de referencia

8. Conclusiones

La metodología propuesta durante este trabajo para el diseño y modelado de un sistema de adquisición de información espectral y de profundidad permitió desarrollar el modelo matemático de adquisición compuesto por estéreo visión con cámaras RGB y una cámara adicional infrarroja. Dicho modelo fue descrito y diseñado específicamente para ser evaluado por una red de inteligencia artificial llamada en este trabajo DNet, que permitió someter a evaluación tanto para entrenamiento como para validación el parámetro libre de distancia entre cámaras, el cual nos permitió definir, que para imágenes de entrenamiento, la configuración que obtiene mejor desempeño en las métricas de devolución PSNR y SSMI fue la de profundidad de 1 metro y distancia entre cámaras de 0.1 metro, además durante esta evaluación, también se observó que a pesar de que la profundidad de 5 metros no fue la ganadora global, se resalta que fue la obtuvo un desempeño mas estable en todas las variaciones de distancia entre cámaras. En cuanto al rendimiento de la red DNet para imágenes de validación, se observa que la distancia óptima entre cámaras para cualquier nivel de profundidad es la distancia de 1 metro. Finalmente, el sistema óptico con los parámetros empelados en el entrenamiento de la red DNet se utilizó para evaluar el rendimiento del sistema real, en un entorno no controlado, con condiciones de iluminación natural, y como resultado se obtuvo que la DNet propuesta obtuvo un desempeño aceptable para implementación.

Referencias Bibliográficas

[Amarsaikhan et al., 2012] Amarsaikhan, D., Saandar, M., Ganzorig, M., Blotevogel, H., Egshiglen, E., Gantuyal, R., Nergui, B., and Enkhjargal, D. (2012). Comparison of multisource image fusion methods and land cover classification. *International Journal of Remote Sensing*, 33(8):2532–2550.

[Camacho et al., 2018] Camacho, A., Vargas, H., and Arguello, H. (2018). Unmixing-based approach as a tool for classification of oil palm diseases using hyperspectral remote sensing in colombia. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XX*, volume 10783, page 1078312. International Society for Optics and Photonics.

[Capper and Elliott, 2013] Capper, P. and Elliott, C. (2013). *Infrared detectors and emitters: materials and devices*, volume 8. Springer Science & Business Media.

[Chen and Blum, 2009] Chen, Y. and Blum, R. S. (2009). A new automated quality assessment algorithm for image fusion. *Image and vision computing*, 27(10):1421–1432.

[Du and Gao, 2017] Du, C. and Gao, S. (2017). Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network. *IEEE access*, 5:15750–15761.

[Fritsch et al., 2013] Fritsch, J., Kuehnl, T., and Geiger, A. (2013). A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*.

[Fu et al., 2019] Fu, J., Gao, X.-R., Xu, M., and Wang, W. (2019). Multi focus and multi-source image fusion based on deep learning model. In *2019 2nd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM)*, pages 512–515.

[Geiger et al., 2013] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.

[Groetsch and Groetsch, 1993] Groetsch, C. W. and Groetsch, C. (1993). *Inverse problems in the mathematical sciences*, volume 52. Springer.

[Kennedy et al., 2007] Kennedy, J. A., Israel, O., Frenkel, A., Bar-Shalom, R., and Azhari, H. (2007). Improved image fusion in pet/ct using hybrid image reconstruction and super-resolution. *International journal of biomedical imaging*, 2007.

[Lakshmipriya et al., 2020] Lakshmipriya, B., Pavithra, N., and Saraswathi, D. (2020). Optimized convolutional neural network based colour image fusion. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–4.

[Lavinia et al., 2016] Lavinia, Y., Vo, H. H., and Verma, A. (2016). Fusion based deep cnn for improved large-scale image action recognition. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 609–614.

[Li et al., 2019] Li, J., Wang, C., Kang, X., and Zhao, Q. (2019). Camera localization for augmented reality and indoor positioning: A vision-based 3d feature database approach. *International journal of digital earth*.

[Li et al., 2013] Li, Y., Wu, H., An, R., Xu, H., He, Q., and Xu, J. (2013). An improved building boundary extraction algorithm based on fusion of optical imagery and lidar data. *Optik*, 124(22):5357–5362.

[Liao et al., 2020] Liao, B., Du, Y., and Yin, X. (2020). Fusion of infrared-visible images in ue-iot for fault point detection based on gan. *IEEE Access*, 8:79754–79763.

[Ma and Karaman, 2018] Ma, F. and Karaman, S. (2018). Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4796–4803. IEEE.

[Malvar et al., 2004] Malvar, H. S., He, L.-w., and Cutler, R. (2004). High-quality linear interpolation for demosaicing of bayer-patterned color images. In *International Conference of Acoustic, Speech and Signal Processing*, pages 485–488. IEEE.

[Martull et al., 2012] Martull, S., Peris, M., and Fukui, K. (2012). Realistic cg stereo image dataset with ground truth disparity maps. In *ICPR workshop TrakMark2012*, volume 111, pages 117–118.

[Menze and Geiger, 2015] Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Park et al., 2018] Park, K., Kim, S., and Sohn, K. (2018). High-precision depth estimation with the 3d lidar and stereo fusion. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2156–2163. IEEE.

[Polyakov et al., 2014] Polyakov, V. M., Vitkin, V. V., Lychagin, D. I., Krylov, A. A., Buchenkov, V. A., and Kashcheev, S. V. (2014). Compact q-switched high repetition rate nd:ylf laser with 100 mj pulse energy for airborne lidars. In *2014 International Conference Laser Optics*, pages 1–2.

[Ramirez and Arguello, 2019] Ramirez, J. M. and Arguello, H. (2019). Multiresolution compressive feature fusion for spectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9900–9911.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

[Scharstein et al., 2014] Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., and Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer.

[Schechner and Kiryati, 2000] Schechner, Y. Y. and Kiryati, N. (2000). Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39(2):141–162.

[Silberman et al., 2012] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer.

[Tsai et al., 1998] Tsai, Y., Leu, J., and Chen, C. (1998). Depth estimation by integration of depth-from-defocus and stereo vision. *Inst. of Information Science, Taipei*.

[Vargas et al., 2019] Vargas, E., Espitia, , Arguello, H., and Tourneret, J.-Y. (2019). Spectral image fusion from compressive measurements. *IEEE Transactions on Image Processing*, 28(5):2271–2282.

[Wang et al., 2020] Wang, C., Du, Q., and Yang, X. (2020). Infrared and visible image fusion method based on vggnet and visual saliency map. In *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*, pages 753–757.

[Wang et al., 2019] Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., and Weinberger, K. Q. (2019). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453.

[Weihua and Shuang, 2019] Weihua, X. and Shuang, F. (2019). Baseline length estimation method based on binocular stereo vision perception. In *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, pages 1120–1123. IEEE.

[Weiss and Biber, 2011] Weiss, U. and Biber, P. (2011). Plant detection and mapping for agricultural robots using a 3d lidar sensor. *Robotics and autonomous systems*, 59(5):265–273.

[Yang et al., 2019] Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., and Zhou, B. (2019). Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 899–908.

[Zhang et al., 2020] Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., and Liu, G. (2020). A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200.

[Zou and Li, 2010] Zou, L. and Li, Y. (2010). A method of stereo vision matching based on opencv. In *2010 International Conference on Audio, Language and Image Processing*, pages 185–190. IEEE.