FUSIÓN DE IMÁGENES DE PROFUNDIDAD OBTENIDAS CON SISTEMAS LIDAR Y DE ESTEREOVISIÓN POR MEDIO DE TÉCNICAS DE APRENDIZAJE PROFUNDO

MIGUEL ANGEL MOLINA GARZÓN
HENRY DARIO MANTILLA CLARO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

FUSIÓN DE IMÁGENES DE PROFUNDIDAD OBTENIDAS CON SISTEMAS LIDAR Y DE ESTEREOVISIÓN POR MEDIO DE TÉCNICAS DE APRENDIZAJE PROFUNDO

MIGUEL ANGEL MOLINA GARZÓN
HENRY DARIO MANTILLA CLARO

Trabajo de Grado para optar al título de
Ingeniero de Sistemas

Director:
Hoover Fabián Rueda-Chacón
Ph.D. en Ingeniería Eléctrica y Computación

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

DEDICATORIA

A mi padre, Henry, por su inquebrantable determinación en forjar un futuro próspero para sus hijos; por sacrificarse incansablemente para que jamás nos falte lo esencial; y por ser el modelo del hombre en el que aspiro convertirme.

A mi madre, Martha, por su amor incondicional y su tierno cariño; por creer en mí, incluso en mis momentos de duda, y enseñarme a levantarme tras cada tropiezo. Gracias por ser el motor que une e impulsa a nuestra familia.

A mis hermanos, Felipe y Valentina, cuyas ocurrencias han llenado mis días de alegría y me han motivado a superarme en cada aspecto, me comprometo a brindarles el mismo apoyo incondicional que recibí de nuestros padres, para que alcancen cada uno de sus sueños.

A mis tías, Rosalba, Ninfa y María, por regalarme un amor incondicional y un apoyo que me ha acompañado a través de los años, y por quererme como a un hijo.

A Linda, por ser mi compañía y apoyo incondicional desde que éramos niños; por siempre creer en mí e impulsarme a desarrollar mi potencial; y por ayudarme a sobrellevar la vida cuando las cosas no salían bien.

A mis abuelas, Isabel y Socorro, por ser el vivo ejemplo de resiliencia y entrega; su sacrificio y esfuerzo para garantizar el bienestar de los suyos han forjado un legado eterno que perpetuará a lo largo del tiempo.

Nada de esto habría sido posible sin ustedes; cada uno de mis triunfos les pertenece.

HENRY DARIO

A Dios, por darme el privilegio de estar viviendo este momento único y especial.

A mi papá Reytembeer, por creer y confiar incondicionalmente en mí, por enseñarme a ser esforzado y valiente, por ser el tipo de hombre que quiero ser y enseñar a mis hijos en un futuro.

A mi mamá Diana Patricia, por enseñarme el valor de la nobleza, por enseñarme a ser valiente aunque tenga temores, y por ser siempre una fuente de amor incondicional.

A mi hermana Maria Fernanda y a mi hermano Juan David, quienes me honran todos los dias con su cariño, con sus sonrisas y con sus ocurrencias, ustedes son el motivo por el cual quiero salir adelante para poder guiarlos y apoyarlos siempre.

A mis abuelas Gloria y Sara, por enseñarme a ser resiliente, generoso y cariñoso sin importar lo dificil que haya sido la vida, que siempre hay un espacio para un tinto y una conversación, esto también es para ustedes.

A mi mascota Paca, por ser esa silenciosa compañía en noches de muchas dudas, por ser una fuente inagotable de amor, y porque su partida me enseñó muchas más cosas de las que pensaba que aprendería.

Por último y no menos importante, quiero hacer una dedicación especial a mi abuelos Gilberto y Alberto, y a mi bisabuela Maria del Pinar. Me encantaría que pudieran leer esto, pero sé que desde el cielo también pueden sentir y compartir mi alegría. Espero poder brindarles muchos y mejores homenajes.

Esto no sería posible sin su amor y su compañía, cada triunfo que cosecho es para ustedes, espero poderles retribuir aunque sea la mitad de lo mucho que me han brindado.

MIGUEL ANGEL

AGRADECIMIENTOS

Agradezco a Dios por todas las bendiciones que me ha otorgado a lo largo de mi vida, por las oportunidades que ha puesto en mi camino

Agradezco a mis padres, los seres mas bondadosos y valientes que he conocido, cuyas virtudes han trascendido a mi vida para moldear quien soy y en quien quiero convertirme.

Agradezco a mi director, Hoover Rueda-Chacón, por depositar su confianza en mí, por su compromiso en el desarrollo de este proyecto y por brindarme la libertad de explorar ideas apasionantes.

Agradezco a mi compañero, Miguel Ángel, por su incansable dedicación al desarrollo de este proyecto y por haber confiado en mis ideas.

Agradezco a mis amigos por cada instante compartido, por las salidas que nos ayudaron a olvidarnos de la rutina y por todas esas veces en que reímos hasta perder el aliento. Cada uno de ustedes es parte esencial de lo que soy. Gracias por todo lo que me han brindado; gracias por todo; espero poder devolverles de corazón todo lo que han hecho por mí.

HENRY DARIO

Agradezco a Dios por darme la vida que tengo, por todas las bendiciones, los regalos y los privilegios que me ha dado desde el primer día que llegué a este mundo.

Agradezco a la universidad pública por recibirme en sus aulas y permitirme desarrollar mis habilidades. Gracias a la Universidad Industrial de Santander por formarme como ingeniero y como persona.

Agradezco a mi director Hoover Rueda-Chacón por su apoyo, su confianza y su guía para desarrollar este proyecto, gracias por motivarme a soñar en grande.

Agradezco a mi compañero Henry Mantilla por su compromiso y su dedicación durante todo este tiempo compartido desarrollando este proyecto.

Agradezco a mis amigos por todos los buenos momentos vividos, por todas las risas, por todos los perros calientes, soy quien soy gracias a ustedes, les deseo lo mejor en sus vidas.

(También gracias a Spotify)

MIGUEL ANGEL

CONTENIDO

LISTA DE FIGURAS

pág.

LISTA DE CUADROS

pág.

RESUMEN

TÍTULO: FUSIÓN DE IMÁGENES DE PROFUNDIDAD OBTENIDAS CON SISTEMAS LIDAR Y DE ESTEREOVISIÓN POR MEDIO DE TÉCNICAS DE APRENDIZAJE PROFUNDO [*]

AUTOR: MIGUEL ANGEL MOLINA GARZÓN, HENRY DARIO MANTILLA CLARO [**]

PALABRAS CLAVE: Algoritmos de fusión, Aprendizaje profundo, Estimación de la profundidad, LiDAR, Estereovisión, *Vision Transformers*.

DESCRIPCIÓN:

La fusión de imágenes de profundidad busca combinar imágenes de profundidad obtenidas mediante modalidades complementarias, como LiDAR y estereovisión, para generar imágenes de profundidad densas y confiables, compensando las limitaciones de los datos LiDAR escasos. Esta tarea resulta crucial en aplicaciones que requieren mediciones precisas de distancia, ya que sensores como los LiDAR, proporcionan mediciones exactas, pero con una cobertura espacial limitada. En contraste, la estereovisión ofrece imágenes de profundidad más densas, aunque con menor precisión en regiones de baja textura o a grandes distancias. En este trabajo, proponemos un método basado en aprendizaje profundo que fusiona la precisión de las mediciones LiDAR con la densidad de puntos que aporta la estereovisión. Para ello, se emplea un modelo preentrenado de correspondencia estéreo que genera imágenes de disparidad, las cuales son refinadas posteriormente mediante una arquitectura basada en *Vision Transformers*, complementada por una red de propagación espacial (SPN). Este enfoque fusiona la información de disparidad con mediciones escasas de profundidad LiDAR, corrigiendo y densificando la imagen de profudidad resultante. Evaluamos nuestro método en el conjunto de datos KITTI, utilizando las métricas RMSE, MAE, iRMSE e iMAE. Nuestro enfoque alcanzó un MAE de $180.88\ mm$, superando a los algoritmos del estado del arte tomados como referencia y obteniendo resultados competitivos en las demás métricas. Nuestro método evidencia la eficacia de utilizar algoritmos pre-entrenados en correspondencia estéreo para generar mapas de profundidad precisos y densos.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Hoover Fabián Rueda-Chacón.

ABSTRACT

TITLE: DEEP LEARNING-BASED FUSION OF LIDAR AND STEREO VISION DEPTH IMAGES [*]

AUTHOR: MIGUEL ANGEL MOLINA GARZÓN, HENRY DARIO MANTILLA CLARO [**]

KEYWORDS: Fusion algorithms, deep learning, depth estimation, LiDAR, stereovision, Vision Transformers.

DESCRIPTION:

Depth image fusion focuses on combining depth images obtained through complementary modalities, such as LiDAR and stereovision, to generate dense and reliable depth images, compensating for the limitations of sparse LiDAR data. This task is crucial in applications that require accurate distance measurements, since sensors like LiDAR provide precise measurements but have limited spatial coverage. In contrast, stereo vision offers denser maps, though with lower precision in low-texture regions or at long distances. In this work, we propose a deep learning-based method that fuses the precision of LiDAR measurements with the point density provided by stereo vision. To do this, we employ a pretrained stereo matching model that generates disparity maps, which are subsequently refined using a Vision Transformer-based architecture, complemented by a Spatial Propagation Network (SPN). This approach fuses disparity information with sparse LiDAR depth measurements, correcting and densifying the resulting map. We evaluate our method on the KITTI dataset using the RMSE, MAE, iRMSE, and iMAE metrics. Our approach achieved a MAE of $180.88\ mm$, outperforming the reference state-of-the-art algorithms and yielding competitive results across the other metrics. Furthermore, our method demonstrates the effectiveness of using pre-trained stereo matching algorithms to produce precise and dense depth maps.

---

[*] Bachelor's Thesis

[**] Faculty of Physical-Mechanical Engineering. School of Systems Engineering & Informatics. Advisor: Hoover Fabián Rueda-Chacón.

INTRODUCCIÓN

El aprendizaje profundo se ha consolidado como un campo fundamental en el ámbito de la inteligencia artificial, destacándose especialmente en tareas de visión por computadora por su capacidad para extraer y modelar patrones complejos en imágenes. En los últimos años arquitecturas innovadoras como los *Vision Transformers (ViTs)* han impulsado el rendimiento en tareas críticas tales como la estimación de la profundidad a partir de sistemas ópticos monoculares[1], la estimación de disparidad[2] y el cálculo del flujo de movimiento[3], gracias a su mecanismo de atención, el cual supera en eficacia a las redes neuronales convolucionales (CNN, del inglés *Convolutional Neural Network*). Estas mejoras han sido especialmente relevantes en el completado de imágenes de profundidad, una tarea que busca estimar las mediciones de profundidad faltantes en datos LiDAR escasos, utilizando la información válida disponible[4]. La precisión en la estimación de los valores de profundidad faltantes es fundamental, ya que permite obtener una representación más completa y precisa de la escena. Esta información es útil para aplicaciones como la navegación autónoma y la robótica, donde es necesario tener una comprensión precisa de la geometría tridimensional del entorno para tomar decisiones correctas en tiempo

---

1 Peng Liu et al. «Transformer-based monocular depth estimation with hybrid attention fusion and progressive regression». En: *Neurocomputing* 620 (2025), pág. 129268.

2 Gangwei Xu et al. «Iterative geometry encoding volume for stereo matching». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 21919-21928.

3 Xiaochen Liu, Tao Zhang y Mingming Liu. «Joint estimation of pose, depth, and optical flow with a competition–cooperation transformer network». En: *Neural Networks* 171 (2024), págs. 263-275.

4 Youmin Zhang et al. «Completionformer: Depth completion with convolutions and vision transformers». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 18527-18536.

real[5]. Las imágenes de profundidad, que representan la distancia de cada píxel en la escena respecto a la cámara, son el resultado de estos cálculos de profundidad.
Para obtener imágenes de profundidad se emplean diversas técnicas, cada una con sus propias ventajas y limitaciones. Por ejemplo, los sistemas de estereovisión utilizan dos cámaras separadas a una distancia fija para adquirir imágenes desde perspectivas distintas, permitiendo la generación de imágenes de profundidad densas mediante técnicas de triangulación[6]. Sin embargo, esta técnica puede presentar inexactitudes en condiciones adversas, tales como áreas con poca textura, regiones afectadas por oclusiones o cuando los objetos se encuentran a grandes distancias. Por el contrario, los sistemas LiDAR (del inglés, *Light Detection And Ranging*) miden la profundidad mediante el cálculo del tiempo de vuelo de pulsos de luz láser, ofreciendo mediciones de profundidad con alta precisión[7]. No obstante, estos datos suelen ser escasos, lo que impide una cobertura densa de la escena.
Para superar estas limitaciones, investigaciones recientes se han orientado hacia el desarrollo de técnicas más robustas, que integran información geométrica tridimensional y que aprovechan modelos fundacionales de inteligencia artificial para completar los mapas de profundidad escasos y mejorar su precisión. La incorporación de información estéreo en el completado de imágenes de profundidad ha demostrado mejorar notablemente el desempeño de las redes neuronales, aunque las técnicas existentes que fusionan dicha información con datos escasos de LiDAR suelen basarse en *encoders* con CNNs, lo que limita su capacidad para capturar

---

5 Lazaros Nalpantidis y Antonios Gasteratos. «Stereo vision for robotic applications in the presence of non-ideal lighting conditions». En: *Image and Vision Computing* 28.6 (2010), págs. 940-951.

6 Heiko Hirschmuller. «Stereo processing by semiglobal matching and mutual information». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (2007), págs. 328-341.

7 You Li y Javier Ibanez-Guzman. «Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems». En: *IEEE Signal Processing Magazine* 37.4 (2020), págs. 50-61.

relaciones globales entre dominios[8]. Esto abre la posibilidad de desarrollar métodos que aprovechen nuevas arquitecturas, como los modelos basados en atención[9], los cuales capturan dependencias globales entre los datos, permitiendo una integración más eficiente y precisa entre distintas modalidades de imágenes de profundidad.
En este trabajo se diseñó un algoritmo de fusión de imágenes de profundidad que combina la precisión de LiDAR con la densidad de la estereovisión, empleando técnicas avanzadas de aprendizaje profundo. Primero se calcula una imagen de disparidad a partir dos imágenes RGB, luego se emplean bloques convolucionales para la extracción de características de una de las imágenes pareadas, junto con la imagen de disparidad y la imagen de profundidad LiDAR. Estas características alimentan un módulo basado en ViTs[10] cuyo objetivo es aprender relaciones entre dichas modalidades. La salidad de este, pasa a un decodificador basado en CNN, que produce una imagen residual de profundidad y una imagen residual de disparidad, un mapa de confidencia y un conjunto de matrices de afinidad. Finalmente, toda esta información se introduce en un módulo de refinamiento[11] lo cual ajusta los valores de profundidad y mejora la coherencia espacial de la imagen generada con respecto a la referencia LiDAR, garantizando una fusión más robusta y precisa en la estimación de profundidad.

---

[8] Ang Li et al. «Stereo-LiDAR Depth Estimation with Deformable Propagation and Learned Disparity-Depth Conversion». En: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2024.

[9] Ashish Vaswani et al. «Attention is All you Need». En: *Advances in Neural Information Processing Systems*. Ed. por I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[10] Wenhai Wang et al. «Pyramid vision transformer: A versatile backbone for dense prediction without convolutions». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 568-578.

[11] Yuankai Lin et al. «Dyspn: Learning dynamic affinity for image-guided depth completion». En: *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

1. OBJETIVOS

Objetivo general

Desarrollar un algoritmo para la fusión de imágenes de profundidad escasas adquiridas con un sistema de detección y medición de distancia por luz (LiDAR) y densas obtenidas mediante estereovisión, utilizando técnicas de aprendizaje profundo, con el objetivo de mejorar la precisión en la estimación de la profundidad de una escena.

Objetivos específicos

1. Identificar, seleccionar y documentar bases de datos adecuadas que contengan imágenes de profundidad escasas obtenidas con un sistema LiDAR junto con imágenes de profundidad densas adquiridas con un sistema de estereovisión.

2. Diseñar un esquema de fusión de imágenes de profundidad basado en algoritmos de aprendizaje profundo, considerando redes neuronales recurrentes, módulos de atención y transformadores de visión.

3. Implementar en Python un algoritmo computacional para mejorar la precisión de imágenes de profundidad adquiridas con un sistema de estereovisión utilizando imágenes de profundidad de un sistema LiDAR y siguiendo el esquema de fusión propuesto.

4. Evaluar el desempeño del algoritmo desarrollado mediante pruebas con las bases de datos disponibles, comparando los resultados con los algoritmos del estado del arte, específicamente *Stereo-LiDAR Depth Estimation with Deformable Propagation and Learned Disparity-Depth Conversion*[8] y *Holistic and*

Contextual Evidential Stereo-LiDAR Fusion for Depth Estimation[12], en términos de métricas de calidad y precisión de mapas de profundidad.

[12] Jiayuan Fan et al. «Holistic and Contextual Evidential Stereo-LiDAR Fusion for Depth Estimation». En: *IEEE Transactions on Intelligent Vehicles* (2024).

2. MARCO DE REFERENCIA

2.1. Imágenes de profundidad

Las imágenes de profundidad, también conocidas como mapas de profundidad, son representaciones visuales en las que a cada píxel se le asigna un valor que indica la distancia desde la cámara hasta el objeto o superficie de la escena en esa posición específica. Este tipo de imágenes proporcionan información crítica de la escena, ya que capturan la geometría espacial en lugar de limitarse a la información de color, como ocurre con las imágenes RGB[13].

Como se puede observar en la Figura 1, las imágenes de profundidad son representaciones de la información tridimensional de una escena, lo que permite una mejor comprensión de la disposición y las relaciones espaciales entre los objetos. Estas imágenes se utilizan ampliamente en tareas como la reconstrucción 3D[14] y la segmentación semántica de objetos[15], ya que añaden una dimensión extra de información comparada con las imágenes RGB tradicionales.

---

[13] Antonio Torralba y Aude Oliva. «Depth estimation from image structure». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.9 (2002), págs. 1226-1238.

[14] Amir Arsalan Soltani et al. «Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks». En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, págs. 1511-1519.

[15] Yanrong Guo y Tao Chen. «Semantic segmentation of RGBD images based on deep depth regression». En: *Pattern Recognition Letters* 109 (2018), págs. 55-64.

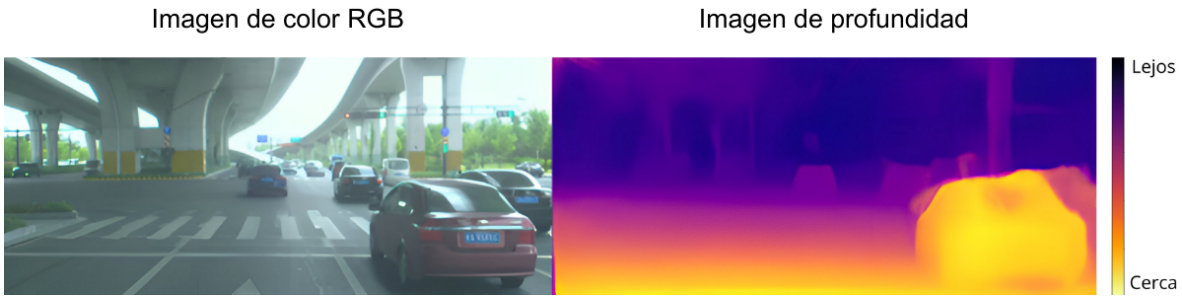

Figura 1. Ejemplo de una imagen de color RGB junto al mapa de profundidad que muestra la distancia de los objetos en una escena. Adaptado de[16].

La precisión y la densidad de la información contenida en una imagen de profundidad pueden variar dependiendo del método de adquisición utilizado. Por lo tanto, la calidad de las imágenes de profundidad está influenciada por varios factores técnicos como la resolución del sensor, la iluminación de la escena, o las propiedades reflectivas de los materiales en la escena[17].

2.2. Sistema de estereovisión

La estereovisión es una técnica de visión por computadora que emula la percepción binocular humana para estimar la profundidad de una escena. Se basa en el uso de dos cámaras, situadas a una distancia conocida entre sí, denominada *baseline*. Una vez rectificadas, al comparar estas imágenes, se pueden calcular las diferencias horizontales entre los puntos correspondientes (disparidad) mediante de métodos de triangulación[6], lo que permite determinar la distancia desde cada punto de la escena a las cámaras. En la Figura 2 se ilustra un sistema de estereovisión donde

---

16 Oscar Real-Moreno et al. «Fast template match algorithm for spatial object detection using a stereo vision system for autonomous navigation». En: *Measurement* 220 (2023), pág. 113299.

17 Jorge Centeno y Boris Jutzi. «Evaluation of a range imaging sensor concerning resolution and illumination». En: *Proceedings, The 2010 Canadian Geomatics Conference and Symposium of Commission I, ISPRS. Calgary, Alberta*. Citeseer. 2010.

dos cámaras adquieren la misma escena desde diferentes perspectivas. El área de superposición entre los campos de visión de las cámaras define la región en la que se puede calcular la disparidad y, por lo tanto, la profundidad.

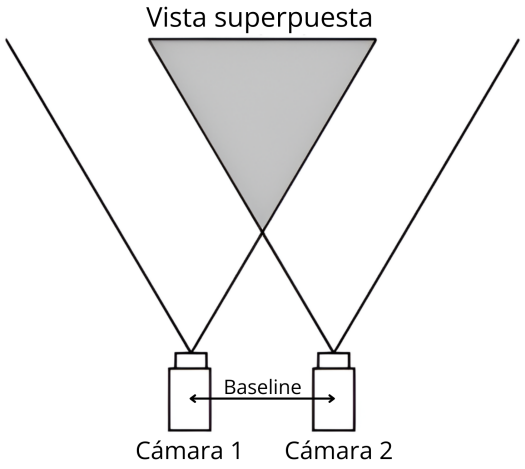

Figura 2. Esquema que ilustra la vista superpuesta obtenida por las dos cámaras en un sistema de estereovisión. Adaptado de[18].

2.2.1. Disparidad

La disparidad (d) describe la diferencia en la posición de un punto observado en las dos imágenes capturadas por las cámaras estéreo (izquierda y derecha). Dado que las cámaras están separadas por una distancia conocida, denominada *baseline* (b), un punto en la escena se proyecta en ubicaciones ligeramente diferentes en cada imagen. La disparidad puede calcularse como:

$$d = x_l - x_r \tag{1}$$

donde:

---

[18] Yasir Mohd Mustafah, Amelia Wong Azman y Fajril Akbar. «Indoor UAV positioning using stereo vision sensor». En: *Procedia Engineering* 41 (2012), págs. 575-579.

- x_l es la coordenada horizontal del punto en la imagen de la cámara izquierda,
- x_r es la coordenada horizontal del mismo punto en la imagen de la cámara derecha.

Este desplazamiento es inversamente proporcional a la distancia del punto respecto a las cámaras[19]. Es decir, los puntos más cercanos tendrán una mayor disparidad, mientras que los más alejados tendrán una menor disparidad. Este principio constituye la base para el cálculo de la profundidad en un sistema estereovisión.

2.2.2. Algoritmos de correspondencia en estereovisión

El cálculo de la disparidad en sistemas de visión estéreo se basa en algoritmos de correspondencia que identifican puntos equivalentes en las imágenes adquiridas por las cámaras izquierda y derecha[20].
Existen diversas estrategias para resolver el problema de correspondencia, que varían en precisión y costo computacional. Estas pueden clasificarse en dos categorías principales:

- **Algoritmos locales:** Comparan regiones (ventanas) de ambas imágenes, utilizando medidas de similitud como la suma de las diferencias cuadráticas (*Sum of Squared Differences, SSD*) o absolutas (*Sum of Absolute Differences, SAD*)[21]. Aunque son eficientes, su precisión disminuye en áreas con baja textura.

---

[19] Ashutosh Saxena, Jamie Schulte, Andrew Y Ng et al. «Depth Estimation Using Monocular and Stereo Cues.» En: *IJCAI*. Vol. 7. 2007, págs. 2197-2203.

[20] Rostam Affendi Hamzah y Haidi Ibrahim. «Literature survey on stereo vision disparity map algorithms». En: *Journal of Sensors* 2016.1 (2016), pág. 8742920.

[21] Karsten Mühlmann et al. «Calculating dense disparity maps from color stereo images, an efficient implementation». En: *International Journal of Computer Vision* 47 (2002), págs. 79-88.

- **Algoritmos globales:** Abordan el problema de correspondencia definiendo un criterio de optimización que se aplica a toda la imagen[6], penalizando grandes variaciones de disparidad para generar imágenes más consistentes.

Entre los métodos existentes existe un compromiso entre precisión y eficiencia computacional, siendo los modelos globales más precisos pero costosos en términos de procesamiento. La elección del algoritmo depende de la aplicación, priorizando velocidad en escenarios de tiempo real y precisión en tareas de reconstrucción detallada.

2.2.3. Triangulación

La triangulación es el principio geométrico que permite calcular la profundidad z de un punto en la escena. Este cálculo se basa en tres elementos fundamentales como, la distancia focal de las cámaras, f , la distancia conocida entre las cámaras (*baseline*), b y la disparidad, d , calculada previamente.
En un sistema estéreo, las cámaras y el punto observado forman un triángulo, como se ilustra en la Figura 3, donde los vértices corresponden al punto en la escena y los lados incluyen el *baseline* y las líneas de proyección desde las cámaras hacia el punto observado.

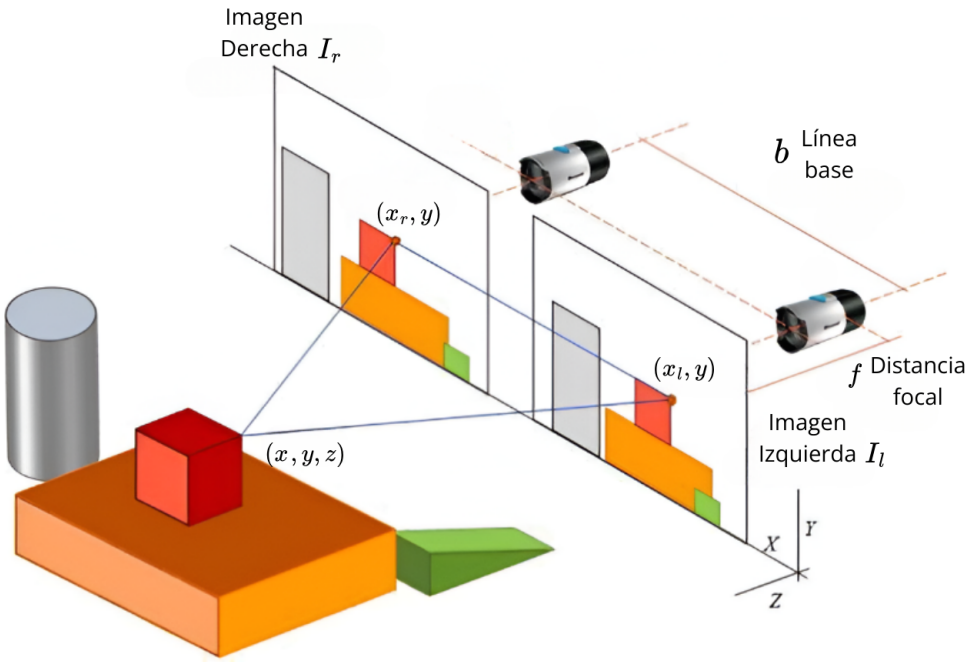

Figura 3. Esquema de un sistema de visión estéreo basado en triangulación, donde en la imagen izquierda (x_l) y derecha (x_r) se presenta la misma escena, captada desde diferentes perspectivas. Adaptado de[22].

Mediante el uso del teorema de semejanza de triángulos, se establece la relación que permite calcular la profundidad z por cada pixel:

$$z = \frac{f \cdot b}{d}. \tag{2}$$

Para llegar a esta ecuación, consideremos que el punto proyectado en la cámara izquierda tiene coordenadas (x_l, y), mientras que en la cámara derecha tiene coordenadas (x_r, y). La relación entre las coordenadas de imagen y las coordenadas tridimensionales (x, y, z) está dada por:

---

[22] Carlos Colodro-Conde et al. «Evaluation of stereo correspondence algorithms and their implementation on FPGA». En: *Journal of Systems Architecture* 60.1 (2014), págs. 22-31.

$$x_l = \frac{x \cdot f}{z}, \quad x_r = \frac{x - b}{z} \cdot f. \tag{3}$$

Restando estas ecuaciones para eliminar z, obtenemos:

$$d = x_l - x_r = \frac{f \cdot b}{z}. \tag{4}$$

Resolviendo para z, se obtiene la ecuación final de profundidad:

$$z = \frac{f \cdot b}{d}. \tag{5}$$

La Ecuación 5 muestra que la profundidad z es inversamente proporcional a la disparidad d. Esto significa que los objetos más cercanos a las cámaras tendrán valores de disparidad mayores, mientras que los objetos más alejados tendrán valores de disparidad menores. Este principio constituye la base para generar imágenes de profundidad que reconstruyen información tridimensional a partir de imágenes estéreo.

2.2.4. Desventajas de estereovisión

Una de las principales ventajas de la estereovisión es su capacidad para generar imágenes de profundidad densas, donde cada píxel de la imagen cuenta con una estimación de la profundidad. Sin embargo, a pesar de su capacidad para producir datos densos, la estereovisión es susceptible a varias fuentes de imprecisión:

1. **Regiones con baja textura**: La estereovisión depende en gran medida de la presencia de características visuales distintivas en las imágenes para emparejar puntos correspondientes. En regiones con baja o nula textura, donde los píxeles tienen valores similares y carecen de detalles diferenciables, el sistema tiene dificultades para determinar con precisión la disparidad, lo que resulta en

errores de estimación de la profundidad[23].

2. **Zonas ocluidas**: Las oclusiones ocurren cuando un objeto bloquea la vista de otro objeto desde una o ambas cámaras. En estas situaciones, la estereovisión no puede emparejar correctamente los puntos entre las dos imágenes, lo que genera áreas de incertidumbre o errores en el mapa de profundidad. Este problema es particularmente relevante en entornos complejos con múltiples objetos dispuestos en diferentes planos[24].

3. **Pérdida de precisión**: A medida que la distancia entre las cámaras y los objetos aumenta, las disparidades entre las imágenes adquiridas por cada cámara disminuyen, volviéndose más difíciles de detectar. Esto reduce la precisión de la estimación de profundidad en objetos lejanos, lo que puede llevar a imágenes de profundidad menos fiables a grandes distancias[25].

Debido a estos factores, aunque la estereovisión proporciona una representación densa de la escena, la precisión de sus imágenes de profundidad puede verse comprometida. En la Figura 4 se muestra un ejemplo de un mapa de profundidad generado por un sistema de estereovisión, donde se puede observar cómo la información de profundidad es densa, pero existen zonas con imprecisiones, especialmente en los bordes y áreas ocluidas.

---

[23] Mingju Chen et al. «Scene reconstruction algorithm for unstructured weak-texture regions based on stereo vision». En: *Applied Sciences* 13.11 (2023), pág. 6407.

[24] Rami Ben-Ari y Nir Sochen. «Variational stereo vision with sharp discontinuities and occlusion handling». En: *2007 IEEE 11th International Conference on Computer Vision*. IEEE. 2007, págs. 1-7.

[25] Manaf A Mahammed, Amera I Melhum y Faris A Kochery. «Object distance measurement by stereo vision». En: *International Journal of Science and Applied Information Technology (IJSAIT)* 2.2 (2013), págs. 05-08.

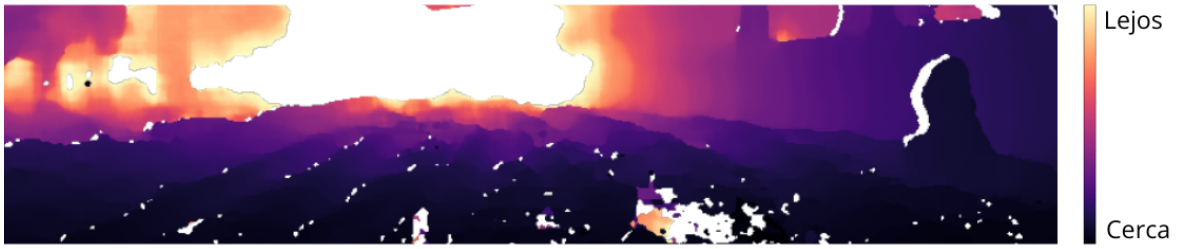

Figura 4. Imágen de profundidad densa obtenida por un algoritmo de correspondencia estéreo, la cual contiene información densa pero imprecisa principalmente en las regiones más alejadas de la cámara y en los bordes de los objetos.

2.3. Sistema LiDAR

LiDAR, del inglés *Light Detection And Ranging*, es una tecnología de medición activa que utiliza pulsos de luz láser para calcular distancias con alta precisión. El principio básico de LiDAR se fundamenta en la emisión de pulsos de luz desde un emisor láser hacia un objeto, y la medición del tiempo que tarda este pulso en regresar al sensor tras ser reflejado por la superficie del objeto. Este tiempo de vuelo, denominado *Time-of-Flight* (Δt), se utiliza para determinar la distancia exacta mediante la ecuación:

$$z = \frac{c \cdot \Delta t}{2} \tag{6}$$

donde: z es la distancia medida, c es la velocidad de la luz en el vacío y Δt es el tiempo que tarda el pulso en realizar el recorrido de ida y vuelta.

LiDAR es considerado un sistema de sensado activo porque emite su propia energía (en forma de pulsos láser) para iluminar los objetos de interés, a diferencia de los sistemas de sensado pasivo que dependen de fuentes de luz externas, como la luz solar o la iluminación ambiental. Esto le permite operar eficazmente en diversas condiciones de iluminación, incluyendo situaciones de baja iluminación o durante la no-

che[26]. La capacidad de LiDAR para medir distancias con alta precisión lo convierte en una herramienta esencial para la construcción de representaciones tridimensionales del entorno, generando un conjunto de datos tridimensionales conocido como "nube de puntos"[27]. En la Figura 5 se ilustra el principio de funcionamiento de un sistema LiDAR típico. El emisor láser proyecta un pulso hacia un objeto, que es reflejado de vuelta al sensor. El tiempo de vuelo es medido para calcular la distancia, representándose gráficamente el proceso junto con un ejemplo de aplicación en un automóvil.

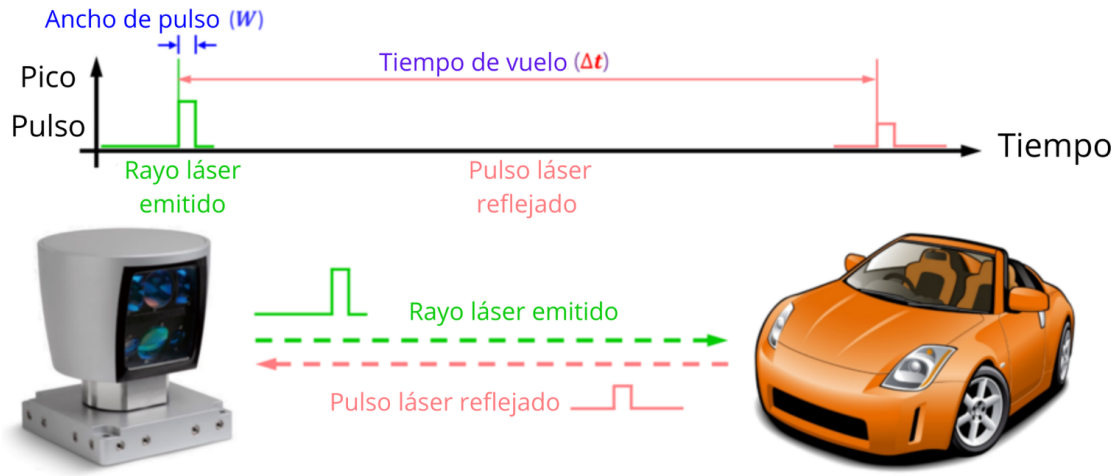

Figura 5. Esquema del principio de funcionamiento: un rayo láser emitido desde el sensor es reflejado por un objeto. El tiempo de vuelo (Δt) entre el pulso emitido y el reflejado se utiliza para calcular con precisión la distancia al vehículo. Adaptado de [28].

---

[26] Xin Wang et al. «The evolution of LiDAR and its application in high precision measurement». En: *IOP Conference Series: Earth and Environmental Science*. Vol. 502. 1. IOP Publishing. 2020, pág. 012008.

[27] Behnam Behroozpour et al. «Lidar system architectures and circuits». En: *IEEE Communications Magazine* 55.10 (2017), págs. 135-142.

[28] Gunzung Kim et al. «Concurrent firing light detection and ranging system for autonomous vehicles». En: *Remote Sensing* 13.9 (2021), pág. 1767.

2.3.1. *Hardware*

El correcto funcionamiento de un sistema LiDAR depende no solo del principio físico de medición basado en el tiempo de vuelo, sino también de la integración y calidad de los componentes de *hardware* que lo componen.

Estos sistemas estan compuestos por un emisor, el cual está acoplado a un foto-detector, lo que permite medir de manera simultánea las distancias hacia múltiples puntos en el entorno, generando una nube de puntos en cada ciclo de rotación[28]. Velodyne LiDAR, una de las empresas pioneras y líderes en la fabricación de sensores LiDAR, revolucionó el mercado con el desarrollo de sistemas multihaz, que permite adquirir datos en un campo de visión de hasta 360° de forma continua, eliminando la necesidad de mover el dispositivo o combinar múltiples sensores para cubrir todo el entorno. Además, la incorporación de múltiples canales láser (por ejemplo, 16, 32 o 64 en modelos como el VLP-16) mejora significativamente la densidad y resolución de los datos recolectados. En la Figura 6 se observa cómo los sensores Velodyne generan nubes de puntos detalladas pero escasas, capaces de representar tanto objetos como superficies en entornos urbanos complejos.

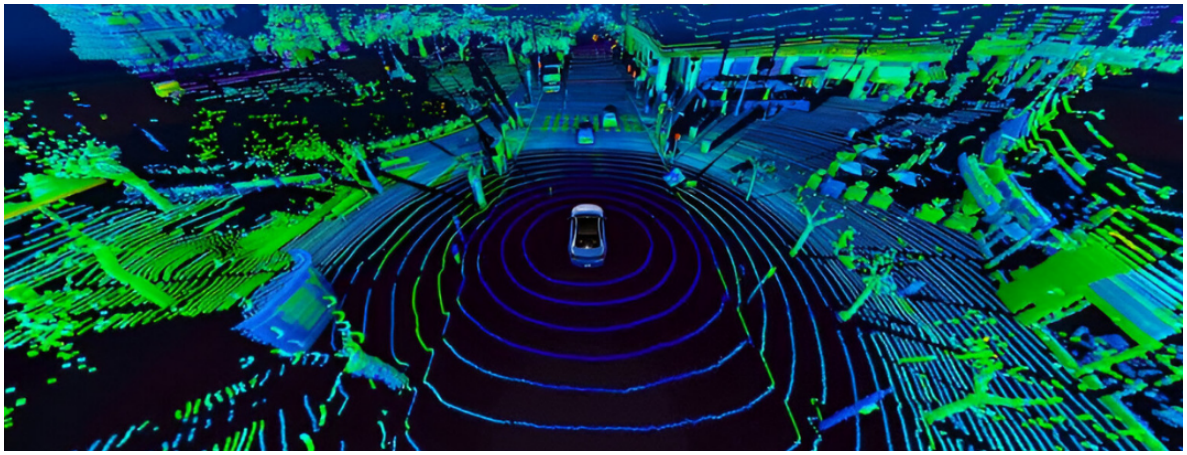

Figura 6. Ejemplo de nube de puntos generada por un sensor Velodyne en un entorno urbano. Tomado de [29].

2.3.2. Desventajas de LiDAR

LiDAR es una tecnología ampliamente utilizada en aplicaciones que requieren datos tridimensionales precisos. Sin embargo, como toda tecnología, presenta tanto ventajas significativas como limitaciones que deben considerarse.

- **Costo elevado:** Los sistemas LiDAR, especialmente aquellos con mayor cantidad de canales (como el Velodyne HDL-64E), pueden ser prohibitivamente costosos, limitando su accesibilidad en proyectos de menor escala [7].
- **Sensibilidad a condiciones atmosféricas:** Factores como la lluvia intensa, la niebla o el polvo pueden degradar la calidad de los datos al dispersar los pulsos de luz láser antes de que lleguen al sensor[26].
- **Procesamiento intensivo:** Manipular y analizar las grandes cantidades de datos generados por un sensor LiDAR requiere *hardware* y *software* avanzados, lo que puede incrementar los costos y la complejidad de los sistemas[30].
- **Datos escasos:** Aunque LiDAR proporciona una precisión geométrica excepcional, la información de profundidad que genera tiende a ser escasa. Los sistemas LiDAR producen una nube de puntos que, aunque es altamente precisa, no cubre toda la superficie de la escena de manera densa. Esto significa que los datos obtenidos a partir de LiDAR son puntuales y dispersos, dejando huecos en la imagen de profundidad. En la Figura 7 se muestra un ejemplo de una imágen de profundidad escasa generada por LiDAR, donde se pueden observar las áreas sin valores en negro.

---

[29] Sam Hind. «Machinic Sensemaking in the Streets: More-than-Lidar in Autonomous Vehicles,'» en: *Seeing the city digitally: Processing urban space and time* (2022), págs. 57-80.

[30] Ievgeniia Maksymova, Christian Steger y Norbert Druml. «Review of LiDAR sensor data acquisition and compression for automotive applications». En: *Eurosensors Conference*. Vol. 2. 13. MDPI. 2018, pág. 852.

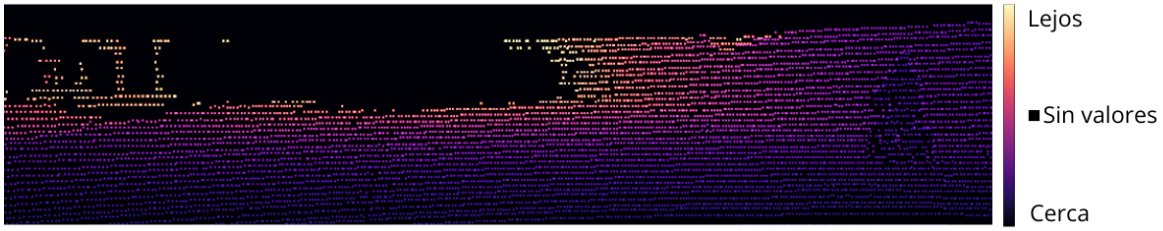

Figura 7. Ejemplo de una imagen de profundidad escasa generada por un sistema LiDAR. Los huecos en la información de profundidad se representan en negro, destacando las áreas sin datos.

2.4. Fusión de imágenes de profundidad

La fusión de imágenes de profundidad es una tarea crucial en visión por computadora que busca combinar múltiples fuentes de datos de profundidad para generar representaciones más completas y precisas. Estas fuentes suelen incluir mediciones escasas obtenidas mediante sistemas de sensado activo, como LiDAR, e imagénes de profundidad generadas a partir de sistemas de estereovisión. La fusión de estas dos modalidades permite aprovechar lo mejor de ambas técnicas, combinando la precisión de LiDAR con la densidad de las imágenes de profundidad generadas por la esterovisión.

A lo largo de los años, se han desarrollado diferentes enfoques para abordar la fusión de imágenes de profundidad. Los métodos clásicos emplean modelos deterministas basados en reglas o algoritmos de interpolación que intentan integrar las mediciones escasas con las densas. Sin embargo, estos enfoques están limitados por su dependencia de parámetros predefinidos y su incapacidad para adaptarse automáticamente a diferentes escenas. Más recientemente, los métodos basados en aprendizaje profundo han revolucionado esta tarea al aprender directamente de los datos. En las siguientes secciones, se explorarán ambos enfoques en detalle.

2.4.1. Métodos clásicos

Los métodos clásicos para la fusión de imágenes de profundidad, también llamada completado de profundidad, se basan en operaciones tradicionales de procesamiento de imágenes y optimización matemática. Estos enfoques, aunque efectivos en ciertos casos, dependen de parámetros predefinidos y no pueden adaptarse automáticamente a diferentes escenarios.

- **Convoluciones básicas:** Un enfoque común consiste en aplicar filtros predefinidos manualmente sobre las imágenes de profundidad escasas para dilatar los valores existentes hacia las regiones vacías[31]. Aunque simples, estos métodos son limitados, ya que no capturan las complejas relaciones espaciales presentes en las escenas.

- **Optimización iterativa:** Algoritmos como la interpolación basada en métodos iterativos ajustan la información de profundidad escasa a través de una función de costo que penaliza las discontinuidades no deseadas[32]. Estos enfoques requieren ajustes manuales de parámetros para cada escena, lo que limita su generalización.

Aunque estos métodos han sido ampliamente utilizados, presentan limitaciones en su capacidad para adaptarse a la variabilidad de los datos y generar resultados precisos en escenarios complejos.

---

[31] Jason Ku, Ali Harakeh y Steven L. Waslander. «In Defense of Classical Image Processing: Fast Depth Completion on the CPU». En: *2018 15th Conference on Computer and Robot Vision (CRV)*. 2018, págs. 16-22. DOI: 10.1109/CRV.2018.00013.

[32] Hongyang Xue, Shengming Zhang y Deng Cai. «Depth image inpainting: Improving low rank matrix completion with low gradient regularization». En: *IEEE Transactions on Image Processing* 26.9 (2017), págs. 4311-4320.

2.4.2. Métodos basados en aprendizaje profundo

En los últimos años, el aprendizaje profundo ha transformado la forma en que se aborda la fusión de imágenes de profundidad. Al aprender directamente de los datos, estos métodos son capaces de capturar patrones complejos y generar imágenes de profundidad densas con mayor precisión. Entre las arquitecturas más utilizadas destacan las *CNNs* y los *ViTs*.

2.4.2.1. Redes Neuronales Convolucionales (CNNs)

Las redes neuronales convolucionales han sido un pilar fundamental en la visión por computadora. Estas redes se componen de capas convolucionales que permiten capturar relaciones espaciales en los datos de entrada, extrayendo características jerárquicas de las imágenes. Los filtros de estas capas convolucionales se ajustan automáticamente mediante *backpropagation*.

Este tipo de red neuronal se caracteriza por su capacidad para modelar relaciones locales en los datos. No obstante, arquitecturas recientes como los ViTs han demostrado un rendimiento superior, atribuyéndose esto a su habilidad para capturar relaciones globales en los datos. La integración de capas convolucionales en arquitecturas ViTs ha demostrado mejorar significativamente su capacidad para representar características locales[33]. En el contexto de la estimación de profundidad a partir de datos LiDAR escasos, las arquitecturas que se basan principalmente en capas convolucionales siguen siendo las más comunes. Para mejorar su rendimiento, se han incorporado módulos de atención basados en convolución[34].

---

[33] Cong Wang et al. «Convolutional embedding makes hierarchical vision transformer stronger». En: *European Conference on Computer Vision*. Springer. 2022, págs. 739-756.

[34] Sanghyun Woo et al. «Cbam: Convolutional block attention module». En: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, págs. 3-19.

2.4.2.2. *Vision Transformers (ViTs)*

Los *transformers* han marcado un hito en el procesamiento del lenguaje natural (NLP) al ofrecer una forma más eficiente, precisa y escalable de capturar relaciones complejas en secuencias de texto, superando las limitaciones de los modelos tradicionales basados en redes neuronales recurrentes. Este éxito ha trascendido el ámbito del lenguaje, inspirando su aplicación en la visión por computadora en tareas tales como clasificación[35], segmentación[36], estimación de la profundidad[37] y completado de profundidad[4].

El desempeño sobresaliente de los *transformers* se fundamenta en el mecanismo de atención, que posibilita capturar relaciones y dependencias entre los elementos de una secuencia[9]. Esto se logra mediante el uso de tres matrices que representan las operaciones fundamentales del mecanismo de atención: consultas (*queries*), claves (*keys*) y valores (*values*) denotadas como $\mathbf{Q}$, $\mathbf{K}$ y $\mathbf{V}$ respectivamente. Estas matrices se generan a partir de transformaciones lineales parametrizadas por matrices ajustables $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$, que mapean los elementos de la secuencia de entrada en las filas correspondientes de las matrices $\mathbf{Q}$, $\mathbf{K}$ y $\mathbf{V}$. El mecanismo de atención calcula un puntaje de similitud, obtenido mediante el producto escalar entre $\mathbf{Q}$ y $\mathbf{K}$. Este puntaje se normaliza y pasa por una función *softmax*, generando un conjunto de pesos de atención que asignan importancia relativa a cada elemento de la secuencia. Finalmente, los valores $\mathbf{V}$ se ponderan con estos pesos, produciendo una

---

[35] Alexey Dosovitskiy et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *International Conference on Learning Representations*. 2021.

[36] Enze Xie et al. «SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers». En: *Neural Information Processing Systems (NeurIPS)*. 2021.

[37] Ashutosh Agarwal y Chetan Arora. «Depthformer: Multiscale Vision Transformer for Monocular Depth Estimation with Global Local Information Fusion». En: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, págs. 3873-3877. DOI: 10.1109/ICIP46576.2022.9897187.

representación que captura las relaciones entre los elementos, así:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{7}$$

En el ámbito de la visión por computadora, se sustituyen las palabras por fragmentos de tamaño fijo de la imagen, denominados parches. Estos parches se transforman en vectores de *embedding* mediante una convolución, a los que posteriormente se añade información posicional a través de otros *embeddings*. Como se ilustra en la Figura 8, este proceso permite emplear el mecanismo de atención para capturar las relaciones y dependencias entre las distintas regiones de la imagen.

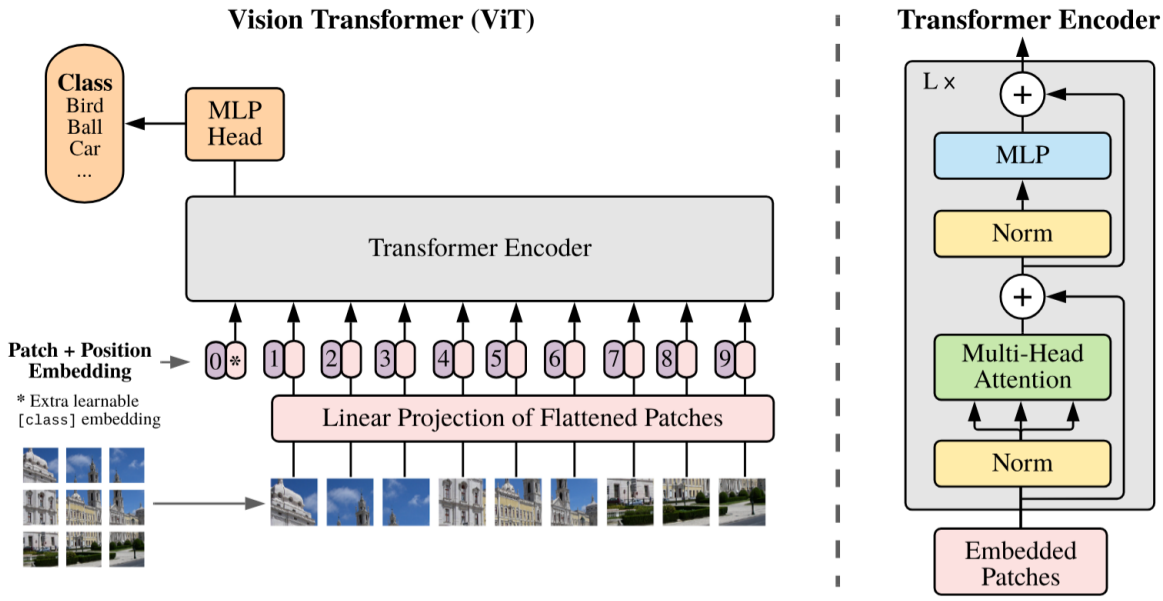

Figura 8. Ilustración de la arquitectura del *encoder* de un *vision transformer* para la tarea de clasificación, mostrando el flujo de información desde la proyección de los parches de imagen y sus codificaciones posicionales en *embeddings*, pasando por el mecanismo de atención multi-cabeza y la red *feed-forward*, hasta capa lineal de clasificación. Adaptado de[35].

3. MÉTODO PROPUESTO

En este trabajo de grado se desarrolló un método para fusionar imágenes de profundidad derivadas tanto de estereovisión como de LiDAR. El modelo se estructura en dos etapas, primero se estima la disparidad entre un par de imágenes RGB mediante un algoritmo de correspondencia estéreo, lo que permite generar una imágen de profundidad densa. La segunda fase, fundamentada en la fusión multimodal, integra los datos de la imágen de profundidad LiDAR con la información de la imagen RGB y la imágen de profundidad densa obtenida previamente mediante la estimación de correspondencia estéreo.

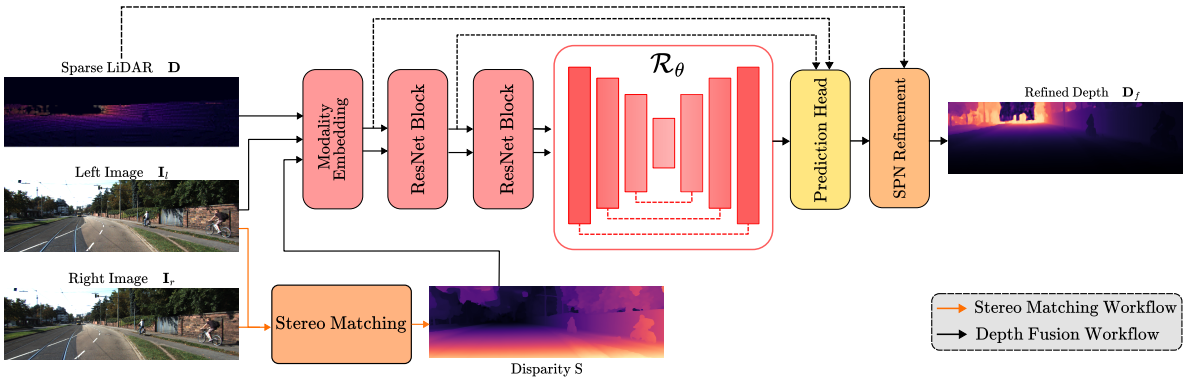

Figura 9. Arquitectura del método propuesto para fusionar imágenes de profundidad $\mathcal{F}_\theta$ obtenidas mediante estereovisión y LiDAR. Inicialmente se obtiene la disparidad mediante un modelo de correspondencia reentrenado en el dataset KITTI-Depth Completion. El sistema emplea bloques convolucionales para extraer características locales. Posteriormente, un *encoder-decoder* basado en ViT procesa dichas características de ambas modalidades, generando representaciones que alimentan la cabeza de predicción. Esta etapa produce imágenes residuales para corregir errores de triangulación en la estereovisión, junto con un mapa de confianza y matrices de afinidad. Finalmente, un módulo de refinamiento integra la información residual, las matrices de afinidad, el mapa de confianza y la imagen LiDAR escasa para generar la imagen de profundidad fusionada.

Como se muestra en la Figura 9, las entradas al modelo de fusión propuesto $\mathcal{F}_\theta$ con-

sisten en un conjunto de imágenes RGB obtenidas mediante un sistema de visión estéreo, representada por $\mathcal{I} = \{\mathbf{I}_l, \mathbf{I}_r\}$, donde $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{H \times W \times 3}$, H y W representan la altura y el ancho de las imágenes, respectivamente, junto con una imagen de profundidad LiDAR escasa $\mathbf{D} \in \mathbb{R}^{H \times W}$. Para estimar la disparidad entre los píxeles de ambas imágenes se emplea un modelo preentrenado de correspondencia estéreo[38] $\mathcal{S}_\theta$. La imagen de disparidad $\mathbf{S} \in \mathbb{R}^{H \times W}$ calculada es convertida a una imagen de profundidad mediante la Ecuación 5; sin embargo, debido a que el error de triangulación crece de manera cuadrática con la distancia, la conversión de disparidad a profundidad produce mediciones imprecisas, principalmente en las regiones más alejadas de la cámara. Con el objetivo de mitigar el error de triangulación, se introduce un algoritmo híbrido $\mathcal{R}_\theta$ que combina bloques convolucionales junto a una arquitectura *transformer encoder-decoder*. Este algoritmo extrae inicialmente características aplicando una capa convolucional a cada entrada ($\mathbf{I}_l$, $\mathbf{D}$, $\mathbf{S}$). Dado que la imagen RGB contiene una mayor variedad de información, como texturas y contrastes, se codifica en un espacio de características de mayor dimensionalidad. En cambio, las imágenes de profundidad y disparidad, al contener información más específica, requieren una representación de menor dimensionalidad. Las características de la imagen $\mathbf{E}_{rgb} \in \mathbb{R}^{H \times W \times 48}$ se concatenan con los extraídas de la imagen de profundidad $\mathbf{E}_p \in \mathbb{R}^{H \times W \times 16}$ y las de la imagen de disparidad $\mathbf{E}_d \in \mathbb{R}^{H \times W \times 16}$. Posteriormente, estas representaciones pasan a través de bloques convolucionales residuales para el dominio de disparidad $\mathcal{C}_d^j$ y profundidad $\mathcal{C}_p^j$, produciendo representaciones $\mathbf{F}_d^i$ y $\mathbf{F}_p^i$ que reducen las dimensiones espaciales a la mitad y aumentan el número de características extraídas, así:

---

[38] Gangwei Xu et al. «IGEV++: Iterative Multi-range Geometry Encoding Volumes for Stereo Matching». En: *arXiv preprint arXiv:2409.00638* (2024).

$$
\begin{aligned}
\mathbf{F}_p^1 &= \mathcal{C}_p^1\left(\mathbf{E}_{rgb} \oplus \mathbf{E}_p\right), \\
\mathbf{F}_d^1 &= \mathcal{C}_d^1\left(\mathbf{E}_{rgb} \oplus \mathbf{E}_d\right), \\
\mathbf{F}_p^2 &= \mathcal{C}_p^2\left(\mathbf{F}_p^1\right), \\
\mathbf{F}_d^2 &= \mathcal{C}_d^2\left(\mathbf{F}_d^1\right),
\end{aligned}
\tag{8}
$$

Donde $\mathbf{F}_p^1, \mathbf{F}_d^1, \mathbf{F}_p^2, \mathbf{F}_d^2 \in \mathbb{R}^{H\times W\times 64}$. El *transformer encoder-decoder* recibe $\mathbf{F}_p$ y $\mathbf{F}_d$, obteniendo como salida del *decoder* una representación $\mathbf{F}_{fusion} \in \mathbb{R}^{H\times W\times 64}$, y a través del módulo de predicción genera dos imágenes residuales, una de disparidad $\mathbf{R}_d \in \mathbb{R}^{H\times W}$ y otra de profundidad $\mathbf{R}_p \in \mathbb{R}^{H\times W}$, una imagen de confidencia $\mathbf{C} \in \mathbb{R}^{H\times W}$ y un conjunto de matrices de afinidad $\{\mathbf{X}^{(k)}\}_{k=1}^{M} \in \mathbb{R}^{H\times W\times M}$, donde M indica el número de matrices de afinidad a generar, así:

$$
\mathbf{R}_d, \mathbf{R}_p, \mathbf{C}, \{\mathbf{X}^{(k)}\}_{k=1}^{M} = \mathcal{R}_\theta\left(\mathbf{F}_p^2, \mathbf{F}_d^2\right). \tag{9}
$$

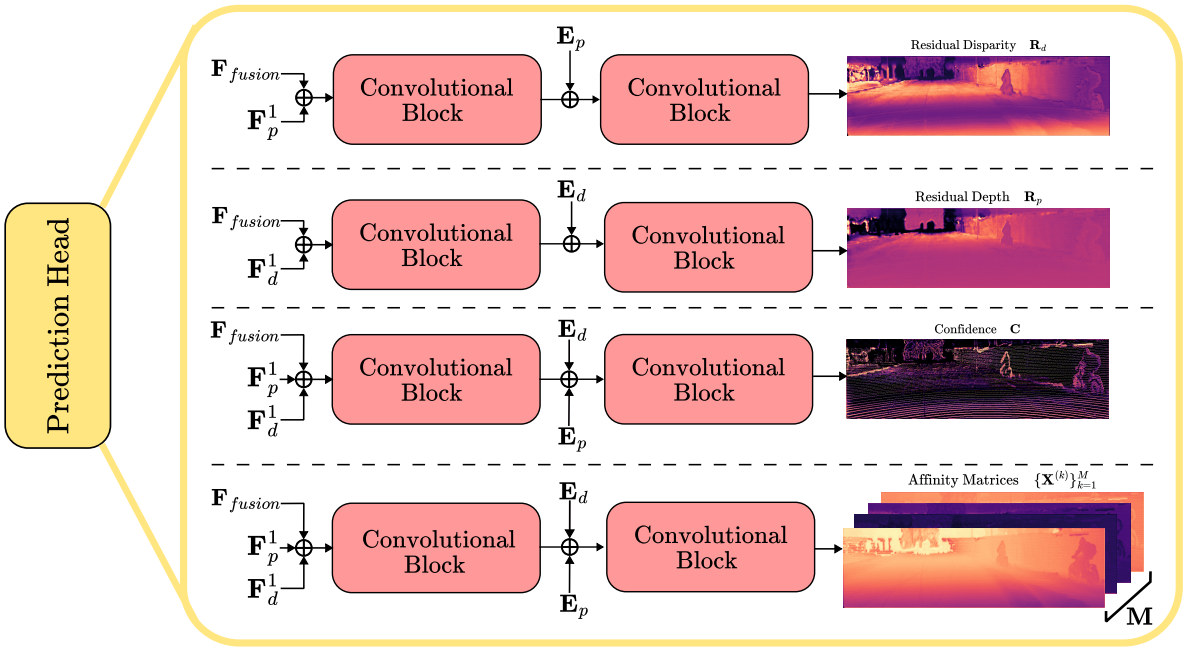

Figura 10. Arquitectura del módulo de predicción que genera las imágenes residuales $\mathbf{R}_p$ y $\mathbf{R}_d$, la imagen de confidencia $\mathbf{C}$ y las matrices de afinidad $\{\mathbf{X}^{(k)}\}_{k=1}^{M}$, para el posterior refinamiento de la imagen de profundidad.

En la Figura 11 se muestra cómo, para aprovechar la relación inversa entre los dominios de disparidad y profundidad, se modificó el diseño del mecanismo de atención con el objetivo de que el modelo pueda establecer correspondencias óptimas entre estos dominios complementarios, además de poder producir una imagen residual de profundidad densa. Primero, se extraen los *embeddings* para ambos dominios, los *embeddings* de profundidad se utilizan para construir la matriz de *queries* $(\mathbf{Q})$ y los *embeddings* de disparidad se usan para construir la matriz de *keys* $(\mathbf{K})$ y *values* $(\mathbf{V})$. De esta manera el mecanismo de atención establece correlaciones precisas entre disparidad y profundidad en las regiones donde existen mediciones LiDAR válidas. Para las regiones donde no existen mediciones LiDAR válidas, se aprovecha de las correlaciones aprendidas para inferir los valores de profundidad faltantes.

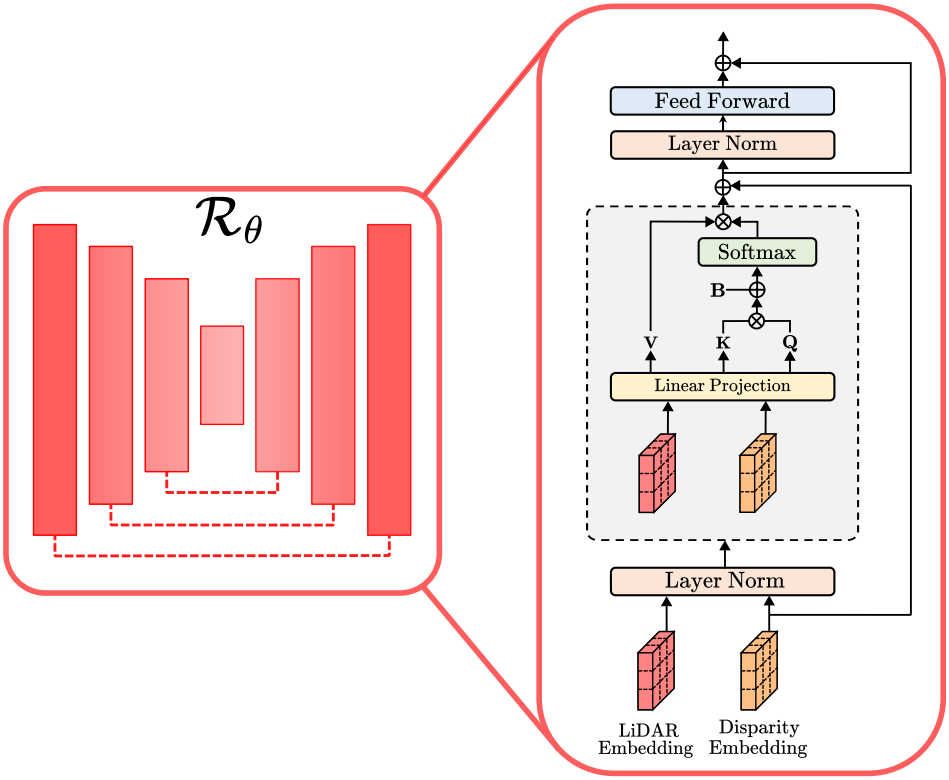

Figura 11. Mecanismo de atención modificado para la fusión de imágenes de profundidad. Adaptado de [39].

Basado en la Ecuación 5, los residuales mitigan los errores de la conversión de la siguiente manera:

$$\mathbf{D}_f = \left(\frac{b \cdot f}{\mathbf{S} + \mathbf{R}_d} \right) + \mathbf{R}_p, \quad \text{donde } \mathbf{R}_d, \mathbf{R}_p \in \mathbb{R}^{H \times W}. \tag{10}$$

Finalmente, la imagen de profundidad densa $\mathbf{D}_f$ y escasa $\mathbf{D}$, la imagen de confidencia $\mathbf{C}$ y las matrices de afinidad $\{\mathbf{X}^{(k)}\}_{k=1}^{M}$ alimentan una red de propagación espacial (SPN, del inglés *Spatial Propagation Network*)[11] que refina $\mathbf{D}_f$ en aquellas regiones donde existen valores de confidencia bajos. Para ello, se usa la imagen de profundidad LiDAR $\mathbf{D}$ como referencia, propagando sus valores sobre $\mathbf{D}_f$ mediante las matrices de afinidad aprendidas. La Figura 12 ilustra el funcionamiento del módulo de refinamiento empleado en nuestro método, identificado como "SPN Refinement" en el diagrama general de la Figura 9.

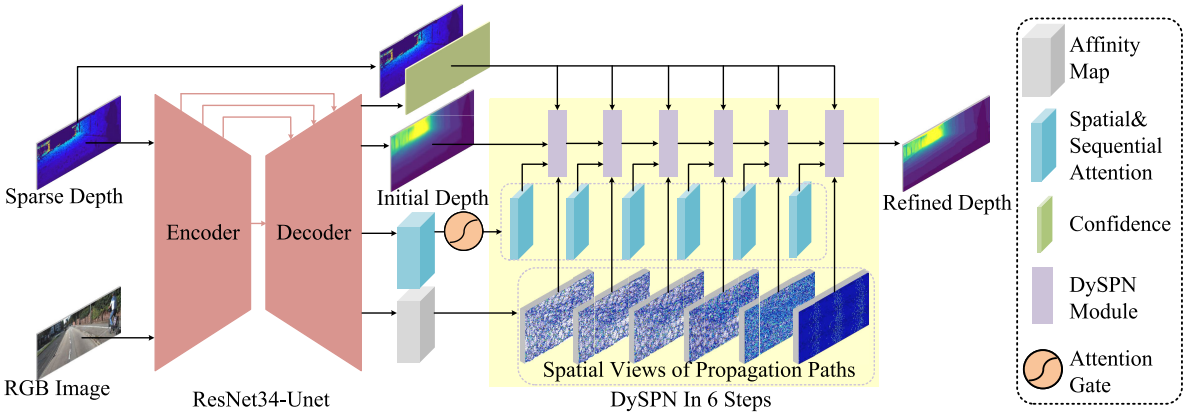

Figura 12. Red de propagación espacial para el refinamiento de imágenes de profundidad. Tomado de [11].

[39] Kyeongha Rho, Jinsung Ha y Youngjung Kim. «Guideformer: Transformers for image guided depth completion». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022, págs. 6250-6259.

4. RESULTADOS

4.1. Base de datos

La evaluación del método propuesto se llevó a cabo utilizando el conjunto de datos *Karlsruhe Institute of Technology and Toyota Technological Institute* (*KITTI*), ampliamente reconocido como un estándar en tareas de percepción para vehículos autónomos[40]. El conjunto de datos KITTI se compone de distintos subconjuntos, cada uno dedicado a tareas específicas como la estimación o el completado de la profundidad[41], la correspondencia estéreo y estimación de flujo[41] y la detección de objetos[42], entre otras. El conjunto de datos KITTI incluye imágenes pareadas RGB, imágenes de profundidad escasas LiDAR junto a su respectivo *ground-truth* semidenso, se define mediante el conjunto $\mathcal{D} = \{\mathbf{I}_l^i, \mathbf{I}_r^i, \mathbf{D}^i, \mathbf{D}_{gt}^i\}_{i=1}^N$ que comprende un total de $N = 42949$ tuplas de imágenes para la partición de entrenamiento y $N = 3426$ para la partición de validación. Las imágenes RGB y las imágenes de profundidad LiDAR cuentan con una resolución espacial de 1242x375 píxeles. Debido a que la parte superior de las imágenes de profundidad no contiene valores válidos, se recortan para obtener una resolución de 1216x256 píxeles durante la inferencia. En la fase de entrenamiento, se utiliza una resolución de 768x256 píxeles para reducir la complejidad computacional. Para obtener el *ground-truth*, se toma la imagen de pro-

---

[40] Jonas Uhrig et al. «Sparsity Invariant CNNs». En: *International Conference on 3D Vision (3DV)*. 2017.

[41] Moritz Menze, Christian Heipke y Andreas Geiger. «Joint 3D Estimation of Vehicles and Scene Flow». En: *ISPRS Workshop on Image Sequence Analysis (ISA)*. 2015.

[42] Andreas Geiger, Philip Lenz y Raquel Urtasun. «Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite». En: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

fundidad escasa LiDAR $\mathbf{D}^i$ como referencia y se acumulan cinco frames anteriores y posteriores a este en una sola imagen de profundidad escasa semi-densa $\mathbf{D}^i_{gt}$ [41]. Como se muestra en la Figura 13, el conjunto de datos incluye imágenes estereoscópicas RGB, imágenes de profundidad generadas a partir de LiDAR y las imágenes de profundidad semi-densas *(ground-truth)*, obtenidas mediante la acumulación de múltiples fotogramas de LiDAR.

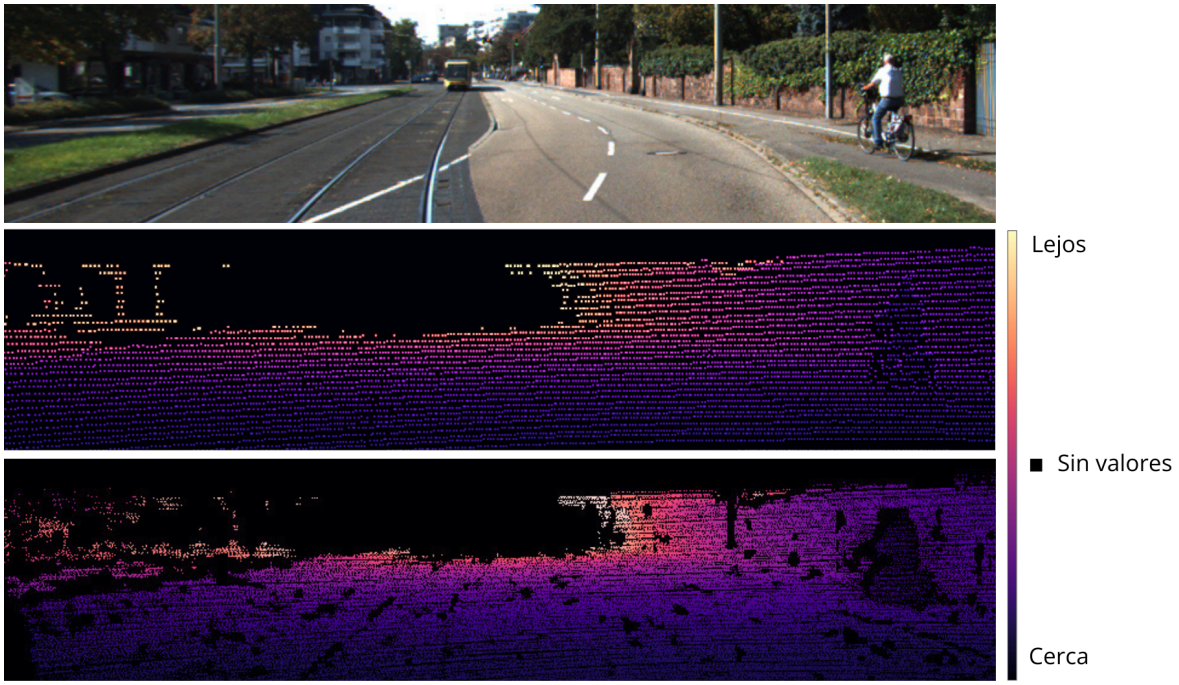

Figura 13. Imagen del conjunto de datos KITTI que muestra una escena urbana RGB (arriba), el mapa de profundidad obtenido con LiDAR (mitad) y la imagen de profundidad semi-densa acumulada (abajo).

4.2. Métricas de evaluación

En la evaluación de algoritmos de fusión de imágenes de estereovisión y LiDAR, se utilizan varias métricas comunes para medir la precisión de los mapas de profundidad generados. Estas métricas permiten cuantificar el error en la estimación de la profundidad y evaluar la efectividad del método propuesto. A continuación, se

describen las métricas más utilizadas en este contexto:

- **Error cuadrático medio inverso (iRMSE)**: El iRMSE, del inglés *Inverse Root Mean Squared Error*, es una métrica que mide el error cuadrático medio de las profundidades inversas en el mapa de profundidad estimado. Se utiliza para evaluar la precisión en la estimación de la profundidad en escenas donde los objetos cercanos tienen mayor importancia, ya que pondera más los errores en las profundidades menores. El iRMSE se expresa en [1/km] y se calcula como:

$$iRMSE = \sqrt{\frac{1}{HW}\sum_{i=1}^{N}\left(\frac{1}{d_i} - \frac{1}{\hat{d}_i}\right)^2}, \tag{11}$$

  donde d_i representa la profundidad real del píxel i, $\hat{d}_i$ es la profundidad estimada para ese mismo píxel y HW es el número total de píxeles.

- **Error absoluto medio inverso (iMAE)**: El iMAE, del inglés *Inverse Mean Absolute Error*, mide el error absoluto medio en las profundidades inversas, similar al iRMSE, pero sin penalizar de manera cuadrática los errores más grandes. Esta métrica es menos sensible a grandes errores puntuales, por lo que es útil cuando se desea evitar la influencia excesiva de puntos fuera de rango en la evaluación del rendimiento. El iMAE se expresa en [1/km]:

$$iMAE = \frac{1}{HW}\sum_{i=1}^{N}\left|\frac{1}{d_i} - \frac{1}{\hat{d}_i}\right|. \tag{12}$$

- **Error Cuadrático Medio (RMSE)**: El RMSE (*Root Mean Squared Error*) es una de las métricas más utilizadas para evaluar la precisión en la estimación de profundidad. Se calcula tomando la raíz cuadrada del promedio de los cuadrados de las diferencias entre la profundidad real y la estimada en cada punto del mapa de profundidad. Esta métrica es especialmente sensible a errores

grandes, lo que permite obtener una medida directa de la magnitud de los errores de estimación. El RMSE se expresa en [mm]:

$$RMSE = \sqrt{\frac{1}{HW}\sum_{i=1}^{N}\left(d_i - \hat{d}_i\right)^2}. \tag{13}$$

- **Error absoluto medio (MAE)**: El MAE, del inglés *Mean Absolute Error*, mide el error absoluto medio entre las profundidades reales y estimadas. A diferencia del RMSE, no penaliza tanto los errores grandes, por lo que puede ser una métrica más adecuada cuando hay outliers o puntos con errores extremadamente grandes. El MAE se expresa en [mm]:

$$MAE = \frac{1}{HW}\sum_{i=1}^{N}\left|d_i - \hat{d}_i\right|. \tag{14}$$

Estas métricas son fundamentales para evaluar la precisión de los algoritmos de fusión de estereovisión y LiDAR, ya que permiten identificar errores en la estimación de profundidad en distintos escenarios y con diferentes niveles de detalle. Utilizar una combinación de estas métricas ayuda a obtener una visión más completa del rendimiento general del sistema de fusión.

4.3. Resultados cuantitativos

Para determinar el desempeño de nuestro algoritmo propuesto, se realizó una comparación con varios métodos del estado del arte ampliamente utilizados en fusión de imágenes de profundidad. Estos métodos representan distintos enfoques en la integración de datos de estereovisión y LiDAR (S+L).

Método	Entradas	RMSE (mm) ↓	MAE (mm) ↓	iRMSE (1/km) ↓	iMAE (1/km) ↓
Listereo [43]	S+L	832.2	283.91	2.190	1.100
GSM [44]	S+L	793.4	271.48	1.531	0.864
CCVN [45]	S+L	749.3	252.50	**1.397**	0.807
S3 [46]	S+L	703.7	239.60	1.540	0.790
SLFNet [47]	S+L	641.1	197.00	1.773	0.876
VPN [48]	S+L	636.2	205.10	1.872	0.987
EG-Depth [49]	S+L	675.5	197.16	1.600	0.787
SDG-Depth[8]	S+L	<u>623.2</u>	197.55	1.519	**0.772**
HCENet[12]	S+L	**599.3**	<u>190.00</u>	<u>1.430</u>	0.780
Nuestro	S+L	625.4	**180.88**	1.604	<u>0.773</u>

Cuadro 1. Comparación de rendimiento para imágenes de profundidad. “S+L” representa estereovisión junto con LiDAR, respectivamente.

---

[43] Junming Zhang et al. «Listereo: Generate dense depth maps from lidar and stereo imagery». En: 2020 *IEEE International Conference on Robotics and Automation (ICRA).* IEEE. 2020, págs. 7829-7836.

[44] Matteo Poggi et al. «Guided stereo matching». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019, págs. 979-988.

[45] Tsun-Hsuan Wang et al. «3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization». En: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE. 2019, págs. 5895-5902.

[46] Yu-Kai Huang et al. «S3: Learnable sparse signal superdensity for guided depth estimation». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, págs. 16706-16716.

[47] Yongjian Zhang et al. «Slfnet: A stereo and lidar fusion network for depth completion». En: *IEEE Robotics and Automation Letters* 7.4 (2022), págs. 10605-10612.

[48] Jaesung Choe et al. «Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation». En: *IEEE Robotics and Automation Letters* 6.3 (2021), págs. 4672-4679.

[49] Zhenyu Xu et al. «Expanding sparse lidar depth and guiding stereo matching for robust dense depth estimation». En: *IEEE Robotics and Automation Letters 8.3* (2023), págs. 1479-1486.

En el Cuadro 1 se presenta una comparación detallada entre “Nuestro” método y diversos enfoques del estado del arte. Cabe destacar que el enfoque desarrollado se caracteriza por una arquitectura más directa y alineada con la naturaleza de los datos, sin la incorporación de mecanismos complejos adicionales. En contraste, HCENet[12] introduce un esquema basado en aprendizaje evidencial para modelar incertidumbres intermodales, lo cual lo hace un enfoque más estadístico. SDG-Depth[8] incorpora módulos sofisticados como la modulación gaussiana del volumen de costos, redes 3D CNN en un esquema *coarse-to-fine*. Si bien estos componentes permiten una estimación detallada y precisa, también hacen que el diseño sea complejo de entender.

En el Cuadro 2 se presenta una comparación porcentual entre las métricas obtenidas por el método propuesto y las reportadas por los dos enfoques más destacados en la tarea de fusión de imágenes de profundidad. Se observa una mejora en la métrica MAE frente a ambos métodos, un rendimiento comparable en RMSE e iMAE, y una desventaja en iRMSE.

Método	**RMSE (mm)** ↓	**MAE (mm)** ↓	**iRMSE (1/km)** ↓	**iMAE (1/km)** ↓
SDG-Depth[8]	0.35 %	8.43 %	5.59 %	0.13 %
HCENet[12]	4.35 %	4.80 %	11.45 %	0.90 %

Cuadro 2. Comparación porcentual del método propuesto frente a SDG-Depth y HCENet. Los valores en verde indican una mejora y los valores en rojo una desmejora con respecto a cada método comparado. La variación porcentual se calcula como la diferencia entre el valor de la métrica obtenida por nuestro método y el valor reportado por el método de referencia, normalizada por este último, esta formulación permite cuantificar el cambio relativo en el rendimiento del método propuesto.

4.4. Estudios de ablación

Se realizó la evaluación del enfoque propuesto en diferentes rangos de profundidad. Esto permite identificar cómo varía el rendimiento del modelo en escenarios cer-

canos, intermedios y lejanos, donde las condiciones de los datos pueden cambiar significativamente.

Rango de profundidad	RMSE (mm) ↓	MAE (mm) ↓	iRMSE (1/km) ↓	iMAE (1/km) ↓
0-20 m	227.2	94.50	1.688	0.871
20-50 m	973.8	408.13	1.170	0.463
50-100 m	2594.7	1148.63	1.030	0.347

Cuadro 3. Comparación de métricas de rendimiento en diferentes rangos de profundidad.

Como se puede observar en el Cuadro 3, las métricas tienden a incrementarse a medida que aumenta la distancia, lo cual es esperado debido a la menor densidad de datos LiDAR y a la menor precisión en la estereovisión a largas distancias. Destaca especialmente el comportamiento en la región lejana (50–100m), donde a pesar del alto RMSE, se logra mantener un iRMSE relativamente bajo lo cual indica una estimación razonablemente precisa de la profundidad en entornos complejos a gran distancia.

Con el fin de evaluar el impacto individual de cada componente del modelo propuesto, se realizaron experimentos ablatorios activando o desactivando módulos clave como el modelo de correspondencia estéreo[38], el módulo de fusión *transformer encoder-decoder (ViT)* , y el módulo de refinamiento[11].

Configuración del modelo			Métricas de rendimiento				# Parámetros (M)
Correspondencia stereo	Fusion ViT	Refinamiento	RMSE (mm) ↓	MAE (mm) ↓	iRMSE (1/km) ↓	iMAE (1/km) ↓	
✓	✗	✗	915.9	340.42	1.740	1.048	14.52
✓	✓	✗	647.6	202.63	1.618	0.858	99.74
✓	✓	✓	625.4	180.88	1.604	0.773	99.74

Cuadro 4. Comparación del uso de módulos, métricas de rendimiento y número de parámetros en distintas configuraciones de modelo.

Los resultados del Cuadro 4 evidencian que cada módulo agregado mejora progresivamente las métricas de rendimiento. En particular, la incorporación del módulo

de fusión genera una mejora significativa respecto al uso exclusivo del modelo de correspondencia estéreo, reduciendo el RMSE en más de $250\ mm$ y el MAE en más de $130\ mm$. Al integrar además el módulo de refinamiento, se obtiene la mejor configuración general, alcanzando el menor MAE e iMAE para el conjunto de datos analizado.

4.5. Resultados cualitativos

Figura 14. Resultados cualitativos del método propuesto. La primera fila presenta la imagen RGB de la izquierda; la segunda, la imagen de disparidad generada por el módulo de *Stereo Matching*; la tercera, la estimación de profundidad resultante de nuestro algoritmo; y la cuarta, el *ground-truth* semi-denso.

En la Figura 14 se muestran los resultados cualitativos del método propuesto. La primera fila presenta la imagen RGB de la izquierda; la segunda, la imagen de disparidad generada por el módulo de *Stereo Matching*; la tercera, la estimación de profundidad obtenida con nuestro método; y la cuarta, el *ground-truth* semi-denso obtenido por LiDAR. Se evidencia que el método propuesto tiene dificultades para generar un mapa de profundidad que conserve los detalles finos de la escena, como los bordes de los objetos.

Figura 15. Visualización de una imagen RGB convertida en una representación tridimensional mediante una nube de puntos, a partir de un mapa de profundidad generado por nuestro método.

En la Figura 15 se muestra cómo, a partir de una imagen RGB y su correspondiente mapa de profundidad, se generó una representación tridimensional de la escena. Para ello, se proyectaron los valores de profundidad sobre la imagen de color utilizando los parámetros intrínsecos de la cámara proporcionados por el conjunto de datos. Esta información se convirtió en una nube de puntos, lo que permite apreciar la precisión de nuestro método para representar la consistencia tridimensional de la escena.

5. CONCLUSIONES

En este trabajo se desarrolló un algoritmo que fusiona imágenes de profundidad precisas pero escasas, obtenidas mediante LiDAR, con imágenes densas aunque menos confiables generadas por estereovisión. El método aprovecha las ventajas de ambas fuentes para estimar imágenes de profundidad densas y precisas, integrando módulos de atención y redes de propagación espacial para el refinamiento de los valores de profundidad estimados.

Los experimentos realizados en el conjunto de datos KITTI demuestran la competitividad de nuestro enfoque en la tarea de fusión de imágenes de profundidad. Nuestro método alcanzó un MAE de $180.88\ mm$, superando a los algoritmos del estado del arte tomados como referencia y obteniendo resultados competitivos en las demás métricas. Además, demostramos la efectividad de emplear algoritmos preentrenados en correspondencia estéreo para complementar los valores escasos de LiDAR, lo que permite que una arquitectura basada en atención establezca relaciones fundamentales entre los dominios y propague de mejor manera los valores de profundidad derivados de la conversión de disparidad a profundidad.

6. TRABAJO FUTURO

Como trabajo futuro, es esencial evaluar la capacidad de generalización del modelo utilizando otros conjuntos de datos de imágenes de profundidad LiDAR, como el dataset MS2[50]. Este dataset integra datos adquiridos mediante LiDAR, estereovisión e incluso sensores NIR, recopilados en diferentes momentos del día y bajo condiciones climáticas variadas, lo que permite explorar enfoques multimodales para la percepción de profundidad más allá de las condiciones diurnas, lo cual resulta de vital importancia en areas como los vehiculos autonomos.

Debido al limitado estudio de este tipo de métodos multimodales, existe una amplia variedad de técnicas que deben ser investigadas para optimizar el rendimiento de esta tarea. Es fundamental explorar la integración de modelos fundacionales como extractores de características, así como implementar mecanismos de atención más eficientes desde el punto de vista computacional. Estas mejoras permitirían reducir la latencia en inferencia e incrementarian el *throughput* del modelo. Estos avances resultan fundamentales para evaluar la viabilidad de modelos híbridos en dispositivos embebidos, donde se demanda una alta tasa de procesamiento por segundo que permita operar en tiempo real sin comprometer significativamente la precisión del modelo.

---

[50] Ukcheol Shin, Jinsun Park e In So Kweon. «Deep Depth Estimation From Thermal Image». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 1043-1053.

BIBLIOGRAFÍA

Agarwal, Ashutosh y Chetan Arora. «Depthformer: Multiscale Vision Transformer for Monocular Depth Estimation with Global Local Information Fusion». En: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, págs. 3873-3877. DOI: 10.1109/ICIP46576.2022.9897187 (vid. pág. 32).

Arsalan Soltani, Amir et al. «Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks». En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, págs. 1511-1519 (vid. pág. 17).

Behroozpour, Behnam et al. «Lidar system architectures and circuits». En: *IEEE Communications Magazine* 55.10 (2017), págs. 135-142 (vid. pág. 26).

Ben-Ari, Rami y Nir Sochen. «Variational stereo vision with sharp discontinuities and occlusion handling». En: *2007 IEEE 11th International Conference on Computer Vision*. IEEE. 2007, págs. 1-7 (vid. pág. 24).

Centeno, Jorge y Boris Jutzi. «Evaluation of a range imaging sensor concerning resolution and illumination». En: *Proceedings, The 2010 Canadian Geomatics Conference and Symposium of Commission I, ISPRS. Calgary, Alberta*. Citeseer. 2010 (vid. pág. 18).

Chen, Mingju et al. «Scene reconstruction algorithm for unstructured weak-texture regions based on stereo vision». En: *Applied Sciences* 13.11 (2023), pág. 6407 (vid. pág. 24).

Choe, Jaesung et al. «Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation». En: *IEEE Robotics and Automation Letters* 6.3 (2021), págs. 4672-4679.

Colodro-Conde, Carlos et al. «Evaluation of stereo correspondence algorithms and their implementation on FPGA». En: *Journal of Systems Architecture* 60.1 (2014), págs. 22-31.

Dosovitskiy, Alexey et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *International Conference on Learning Representations*. 2021 (vid. pág. 32).

Fan, Jiayuan et al. «Holistic and Contextual Evidential Stereo-LiDAR Fusion for Depth Estimation». En: *IEEE Transactions on Intelligent Vehicles* (2024) (vid. pág. 16).

Geiger, Andreas, Philip Lenz y Raquel Urtasun. «Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite». En: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012 (vid. pág. 39).

Guo, Yanrong y Tao Chen. «Semantic segmentation of RGBD images based on deep depth regression». En: *Pattern Recognition Letters* 109 (2018), págs. 55-64 (vid. pág. 17).

Hamzah, Rostam Affendi y Haidi Ibrahim. «Literature survey on stereo vision disparity map algorithms». En: *Journal of Sensors* 2016.1 (2016), pág. 8742920 (vid. pág. 20).

Hind, Sam. «Machinic Sensemaking in the Streets: More-than-Lidar in Autonomous Vehicles,'» en: *Seeing the City Digitally: Processing Urban Space and Time* (2022), págs. 57-80.

Hirschmuller, Heiko. «Stereo processing by semiglobal matching and mutual information». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (2007), págs. 328-341 (vid. pág. 13).

Huang, Yu-Kai et al. «S3: Learnable sparse signal superdensity for guided depth estimation». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, págs. 16706-16716.

Kim, Gunzung et al. «Concurrent firing light detection and ranging system for autonomous vehicles». En: *Remote Sensing* 13.9 (2021), pág. 1767.

Ku, Jason, Ali Harakeh y Steven L. Waslander. «In Defense of Classical Image Processing: Fast Depth Completion on the CPU». En: *2018 15th Conference on Computer and Robot Vision (CRV)*. 2018, págs. 16-22. DOI: 10.1109/CRV.2018.00013 (vid. pág. 30).

Li, Ang et al. «Stereo-LiDAR Depth Estimation with Deformable Propagation and Learned Disparity-Depth Conversion». En: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2024 (vid. pág. 14).

Li, You y Javier Ibanez-Guzman. «Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems». En: *IEEE Signal Processing Magazine* 37.4 (2020), págs. 50-61 (vid. pág. 13).

Lin, Yuankai et al. «Dyspn: Learning dynamic affinity for image-guided depth completion». En: *IEEE Transactions on Circuits and Systems for Video Technology* (2023) (vid. pág. 14).

Liu, Peng et al. «Transformer-based monocular depth estimation with hybrid attention fusion and progressive regression». En: *Neurocomputing* 620 (2025), pág. 129268 (vid. pág. 12).

Liu, Xiaochen, Tao Zhang y Mingming Liu. «Joint estimation of pose, depth, and optical flow with a competition–cooperation transformer network». En: *Neural Networks* 171 (2024), págs. 263-275 (vid. pág. 12).

Mahammed, Manaf A, Amera I Melhum y Faris A Kochery. «Object distance measurement by stereo vision». En: *International Journal of Science and Applied Information Technology (IJSAIT)* 2.2 (2013), págs. 05-08 (vid. pág. 24).

Maksymova, Ievgeniia, Christian Steger y Norbert Druml. «Review of LiDAR sensor data acquisition and compression for automotive applications». En: *Eurosensors Conference*. Vol. 2. 13. MDPI. 2018, pág. 852 (vid. pág. 28).

Menze, Moritz, Christian Heipke y Andreas Geiger. «Joint 3D Estimation of Vehicles and Scene Flow». En: *ISPRS Workshop on Image Sequence Analysis (ISA)*. 2015 (vid. pág. 39).

Mühlmann, Karsten et al. «Calculating dense disparity maps from color stereo images, an efficient implementation». En: *International Journal of Computer Vision* 47 (2002), págs. 79-88 (vid. pág. 20).

Mustafah, Yasir Mohd, Amelia Wong Azman y Fajril Akbar. «Indoor UAV positioning using stereo vision sensor». En: *Procedia Engineering* 41 (2012), págs. 575-579.

Nalpantidis, Lazaros y Antonios Gasteratos. «Stereo vision for robotic applications in the presence of non-ideal lighting conditions». En: *Image and Vision Computing* 28.6 (2010), págs. 940-951 (vid. pág. 13).

Poggi, Matteo et al. «Guided stereo matching». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, págs. 979-988.

Real-Moreno, Oscar et al. «Fast template match algorithm for spatial object detection using a stereo vision system for autonomous navigation». En: *Measurement* 220 (2023), pág. 113299.

Rho, Kyeongha, Jinsung Ha y Youngjung Kim. «Guideformer: Transformers for image guided depth completion». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 6250-6259.

Saxena, Ashutosh, Jamie Schulte, Andrew Y Ng et al. «Depth Estimation Using Monocular and Stereo Cues.» En: *IJCAI*. Vol. 7. 2007, págs. 2197-2203 (vid. pág. 20).

Shin, Ukcheol, Jinsun Park e In So Kweon. «Deep Depth Estimation From Thermal Image». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 1043-1053 (vid. pág. 49).

Torralba, Antonio y Aude Oliva. «Depth estimation from image structure». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.9 (2002), págs. 1226-1238 (vid. pág. 17).

Uhrig, Jonas et al. «Sparsity Invariant CNNs». En: *International Conference on 3D Vision (3DV)*. 2017 (vid. pág. 39).

Vaswani, Ashish et al. «Attention is All you Need». En: *Advances in Neural Information Processing Systems*. Ed. por I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017 (vid. pág. 14).

Wang, Cong et al. «Convolutional embedding makes hierarchical vision transformer stronger». En: *European Conference on Computer Vision*. Springer. 2022, págs. 739-756 (vid. pág. 31).

Wang, Tsun-Hsuan et al. «3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization». En: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, págs. 5895-5902.

Wang, Wenhai et al. «Pyramid vision transformer: A versatile backbone for dense prediction without convolutions». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 568-578 (vid. pág. 14).

Wang, Xin et al. «The evolution of LiDAR and its application in high precision measurement». En: *IOP Conference Series: Earth and Environmental Science*. Vol. 502. 1. IOP Publishing. 2020, pág. 012008 (vid. pág. 26).

Woo, Sanghyun et al. «Cbam: Convolutional block attention module». En: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, págs. 3-19 (vid. pág. 31).

Xie, Enze et al. «SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers». En: *Neural Information Processing Systems (NeurIPS)*. 2021 (vid. pág. 32).

Xu, Gangwei et al. «IGEV++: Iterative Multi-range Geometry Encoding Volumes for Stereo Matching». En: *arXiv preprint arXiv:2409.00638* (2024) (vid. pág. 35).

Xu, Gangwei et al. «Iterative geometry encoding volume for stereo matching». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 21919-21928 (vid. pág. 12).

Xu, Zhenyu et al. «Expanding sparse lidar depth and guiding stereo matching for robust dense depth estimation». En: *IEEE Robotics and Automation Letters* 8.3 (2023), págs. 1479-1486.

Xue, Hongyang, Shengming Zhang y Deng Cai. «Depth image inpainting: Improving low rank matrix completion with low gradient regularization». En: *IEEE Transactions on Image Processing* 26.9 (2017), págs. 4311-4320 (vid. pág. 30).

Zhang, Junming et al. «Listereo: Generate dense depth maps from lidar and stereo imagery». En: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, págs. 7829-7836.

Zhang, Yongjian et al. «Slfnet: A stereo and lidar fusion network for depth completion». En: *IEEE Robotics and Automation Letters* 7.4 (2022), págs. 10605-10612.

Zhang, Youmin et al. «Completionformer: Depth completion with convolutions and vision transformers». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 18527-18536 (vid. pág. 12).