# ESTIMACIÓN PASIVA DE LA PROFUNDIDAD A PARTIR DE IMÁGENES HIPERESPECTRALES EN EL INFRARROJO DE ONDA LARGA (LWIR) MEDIANTE APRENDIZAJE PROFUNDO GUIADO POR LA FÍSICA

GUILLERMO PINTO RUIZ

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

# ESTIMACIÓN PASIVA DE LA PROFUNDIDAD A PARTIR DE IMÁGENES HIPERESPECTRALES EN EL INFRARROJO DE ONDA LARGA (LWIR) MEDIANTE APRENDIZAJE PROFUNDO GUIADO POR LA FÍSICA

GUILLERMO PINTO RUIZ

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 abuela Fidelia. Me enseñaste que se puede salir adelante a pesar de cualquier circunstancia.

A mi madre, Adela, por levantarse todos los días a darme ánimos para continuar, tú eres la verdadera guerrera.

A mi hermana, Sara, por sacarme de la rutina y regalarme un abrazo cada vez que la veo.

A mi padre, Guillermo, por apoyarme sin juzgar el porqué de mis acciones.

A la gata de mi hermana, Mia, por acompañarme silenciosamente los fines de semana.

**GUILLERMO PINTO**

## AGRADECIMIENTOS

Agradezco a mis padres, por haberme mostrado que para salir adelante hay que trabajar muy duro. Y por todo el apoyo que me han dado para seguir mi camino.

A mi director, Hoover, por haber creído en mi, por todo el tiempo que me ha dedicado y la libertad que me ha dado para explorar ideas.

A los chicos del semillero Hands-on Computer Vision. Han sido mis únicos compañeros en el último año. Todos tienen un potencial inmenso, solo deben creérselo.

A mi gran amiga, Nohelia, por acompañarme en varios de mis logros y fracasos. Te deseo lo mejor siempre.

Por último, pero no menos importante, gracias a mi, por creer en mi.

**GUILLERMO PINTO**

**CONTENIDO**


**pág.**

**INTRODUCCIÓN** **12**

**1 OBJETIVOS** **16**

**2 MARCO DE REFERENCIA** **17**

2.1 Estimación pasiva de la profundidad 17

2.2 Radiación infrarroja 20

2.3 Imágenes infrarrojas hiperespectrales 23

2.4 Aprendizaje profundo guiado por la física 26

**3 MÉTODO PROPUESTO** **31**

3.1 Modelo de formación de imagen 32

3.2 Adaptación del *Transformer* pre-entrenado 35

3.3 Integración de los *decoders* de profundidad, temperatura y emisividad 37

3.4 Planteamiento de la función de pérdida física 40

**4 RESULTADOS** **42**

4.1 Base de datos 42

4.1.1 Procedimiento de simulación 43

4.1.2 Partición de la base de datos 45

4.2 Métricas de evaluación 47

4.3 Simulaciones 49

4.3.1 Estudios de ablación 49

4.3.2 Resultados cuantitativos 54

4.3.3 Resultados cualitativos 57

4.3.4 Resultados experimentales 59

**5 CONCLUSIONES 68**

**6 TRABAJO FUTURO 69**

**BIBLIOGRAFÍA 70**

## LISTA DE FIGURAS

**pág.**


Figura 1 Ejemplo comparativo de profundidad relativa y absoluta. 18

Figura 2 Simulación de medidas hiperespectrales (en microflicks) de la emisión de una roca a 300 K, registrada a 10, 20 y 30 m. 20

Figura 3 Espectro de radiación para cinco cuerpos negros con temperaturas de 100, 500, 1000, 3000 y 10000 K. 22

Figura 4 Comparación esquemática de las imágenes multiespectrales e hiperespectrales. 24

Figura 5 Flujo de procesamiento de un ViT. 28

Figura 6 Esquema del enfoque propuesto en PUDE, donde las funciones de pérdida se derivan a partir de estimaciones generadas por un modelo de referencia entrenado en aire (DPT). 30

Figura 7 Esquema general de la arquitectura propuesta para la estimación de la profundidad a partir de imágenes hiperespectrales en el LWIR. 32

Figura 8 Diagrama de la capa `HyperspectralPatchEmbed`, empleada para adaptar la proyección inicial del *Transformer* pre-entrenado a imágenes hiperespectrales. 37

Figura 9 Mapas de temperatura, profundidad y emisividad para el primer fotograma de varias escenas sintéticas de la base de datos *HADAR Database*. 44

Figura 10 Relación entre emisividades y funciones de atenuación. 46

Figura 11 Curvas de pérdida en la etapa de entrenamiento para D(ET)-H con $\phi = 1$ y $\phi = 15$. 54

Figura 12 Comparación cualitativa de los resultados de estimación de la profundidad con entradas pseudo-RGB (*Sum* y PCA). 58
Figura 13 Comparación cualitativa de los resultados de estimación de la profundidad al emplear directamente la entrada hiperespectral (HSI). 59
Figura 14 Imágenes pseudo-broadband y etiquetas de profundidad obtenidas para algunas escenas de la base de datos *IH Dataset*. 62
Figura 15 Comparación cualitativa de los resultados de estimación de la profundidad en una escena real de la base de datos *IH Dataset* frente al método empleado en HADAR. 65
Figura 16 Comparación cualitativa de los resultados de estimación de la profundidad en una escena real de la base de datos *IH Dataset* frente al método empleado en *Absorption-Based, Passive Range Imaging From Hyperspectral Thermal Measurements*. 67

## LISTA DE CUADROS

**pág.**

Cuadro 1 Resultados del estudio de ablación para la estimación de la profundidad. 53
Cuadro 2 Comparación con enfoques de referencia para la estimación de la profundidad en la base de datos sintética. 56
Cuadro 3 Comparación con enfoques de referencia del estado del arte para la estimación de la profundidad en la base de datos *IH Dataset*. 63

# RESUMEN

**TÍTULO:** ESTIMACIÓN PASIVA DE LA PROFUNDIDAD A PARTIR DE IMÁGENES HIPERESPECTRALES EN EL INFRARROJO DE ONDA LARGA (LWIR) MEDIANTE APRENDIZAJE PROFUNDO GUIADO POR LA FÍSICA[*]

**AUTOR:** GUILLERMO PINTO RUIZ[**]

**PALABRAS CLAVE:** Aprendizaje profundo, Estimación de la profundidad, Imágenes térmicas, Imágenes hiperespectrales.

**DESCRIPCIÓN:**

La estimación pasiva de la profundidad constituye un reto fundamental en la visión por computadora, especialmente en escenarios con baja visibilidad donde los enfoques basados en el espectro visible fallan. La radiación térmica en el infrarrojo de onda larga (LWIR) ofrece una alternativa prometedora, pues aprovecha la emisión natural de calor de los objetos, permitiendo su observación sin necesidad de iluminación activa. En este trabajo presentamos una arquitectura de aprendizaje profundo para la estimación de la profundidad a partir de imágenes hiperespectrales en el LWIR, integrada con un modelo físico de formación de imagen que guía el entrenamiento. El método propuesto combina un *Transformer encoder* pre-entrenado, adaptado a imágenes hiperespectrales, con tres *decoders* para estimar los mapas de profundidad, temperatura y emisividad. Estas predicciones se guían por una función de pérdida física basada en el modelo de formación de imagen, con la intención de que tengan coherencia con las propiedades reales de la escena. Evaluamos exhaustivamente nuestro método en un conjunto de datos sintético y realizamos validaciones en escenarios reales. Los resultados demuestran que la propuesta es comparable a los métodos de referencia, mostrando una estimación de la profundidad que preserva detalles finos y estructuras complejas en escenarios sintéticos y mejor capacidad de generalización en escenas del mundo real.

---

[*] Trabajo de grado

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

## ABSTRACT

**TITLE:** PASSIVE RANGING FROM LONG-WAVE INFRARED (LWIR) HYPERSPECTRAL IMAGERY USING PHYSICS-GUIDED DEEP LEARNING[*]

**AUTHOR:** GUILLERMO PINTO RUIZ[**]

**KEYWORDS:** Deep Learning, Passive Ranging, Thermal Imaging, Hyperspectral Imaging.

**DESCRIPTION:**

Passive depth estimation is a fundamental challenge in computer vision, particularly in low-visibility scenarios where visible-spectrum approaches fail. Long-wave infrared (LWIR) thermal radiation offers a promising alternative, as it leverages the natural heat emission of objects, enabling observation without active illumination. In this work, we present a deep learning architecture for depth estimation from LWIR hyperspectral imagery, integrated with a physics-based image formation model that guides training. The proposed method combines a pretrained Transformer encoder, adapted to hyperspectral imagery, with three decoders to estimate depth, temperature, and emissivity maps. These predictions are constrained by a physics-inspired loss function based on the image formation model, ensuring consistency with the physical properties of the scene. We thoroughly evaluate our approach on a synthetic dataset and conduct validations in real-world scenarios. The results demonstrate that the proposed method is comparable to state-of-the-art approaches, showing depth estimation that preserves fine details and complex structures in synthetic scenarios and better generalization capabilities in real-world scenes.


---

[*] Bachelor's Thesis

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

# INTRODUCCIÓN

La inteligencia artificial y, en particular, el aprendizaje profundo han transformado radicalmente la forma en que los sistemas inteligentes perciben e interpretan su entorno, impulsando avances notables en áreas como la visión por computadora, la robótica y la percepción remota.[1,2] Estos desarrollos han permitido la automatización de tareas complejas y la adopción de enfoques cada vez más sofisticados para la interpretación de escenas visuales en condiciones diversas.[3] En este contexto, la estimación de la profundidad a partir de imágenes se ha consolidado como una capacidad fundamental para aplicaciones que requieren reconstrucción tridimensional,[4] navegación autónoma[5] y el análisis preciso de escenarios,[6] especialmente en ambientes donde la visibilidad es limitada.[7] El uso de imágenes hiperespectrales en el infrarrojo de onda larga (LWIR, del inglés, *LongWave InfraRed*) representa una

---

1 Yann LeCun, Yoshua Bengio y Geoffrey Hinton. «Deep learning». En: *Nature* 521.7553 (2015), págs. 436-444.

2 Raia Hadsell et al. «Learning long-range vision for autonomous off-road driving». En: *Journal of Field Robotics* 26.2 (2009), págs. 120-144.

3 Richard Szeliski. *Computer vision: Algorithms and applications*. Springer Nature, 2022.

4 Ashutosh Saxena, Sung H Chung y Andrew Y Ng. «3-d depth reconstruction from a single still image». En: *International Journal of Computer Vision* 76 (2008), págs. 53-69.

5 Andreas Geiger, Philip Lenz y Raquel Urtasun. «Are we ready for autonomous driving? the kitti vision benchmark suite». En: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, págs. 3354-3361.

6 Haotong Lin et al. «Prompting Depth Anything for 4K Resolution Accurate Metric Depth Estimation». En: *arXiv preprint arXiv:2412.14015* (2024).

7 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.

alternativa prometedora frente a las soluciones tradicionales, ya que permite la percepción pasiva de escenas mediante la captación de la radiación térmica emitida naturalmente por los objetos.[8]

Sin embargo, la recuperación precisa de información física a partir de estas imágenes enfrenta retos considerables. La señal registrada por el sensor depende simultáneamente de la distancia de los objetos al sensor, su temperatura, su emisividad, su reflectancia y la interacción con la atmósfera, lo que dificulta la descomposición de los atributos físicos relevantes y limita la capacidad de estimar la profundidad de manera confiable.[9] Abordar este desafío es crucial, ya que una estimación pasiva y precisa de la profundidad abriría nuevas posibilidades para sistemas de percepción en entornos donde el uso de sensores activos es impráctico o costoso,[10,11] y donde la interpretación detallada de la escena puede marcar la diferencia en la toma de decisiones autónomas.[12]

Diversos trabajos recientes han intentado superar estas limitaciones mediante la incorporación de modelos físicos y técnicas de aprendizaje profundo. Un ejemplo de ello es el desarrollo de técnicas de estimación pasiva de la profundidad, la temperatura y la emisividad, a partir de mediciones multiespectrales en el LWIR, donde se

---

[8] Yasuto Nagase et al. «Shape from thermal radiation: Passive ranging using multi-spectral lwir measurements». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 12661-12671.

[9] Dimitris G Manolakis, Ronald B Lockwood y Thomas W Cooley. *Hyperspectral imaging remote sensing: Physics, sensors, and algorithms*. Cambridge University Press, 2016.

[10] Zhexuan Cao et al. «Aberration-robust monocular passive depth sensing using a meta-imaging camera». En: *Light: Science & Applications* 13.1 (2024), pág. 236.

[11] George M Williams Jr. «Optimization of eyesafe avalanche photodiode lidar for automobile safety and autonomous navigation systems». En: *Optical Engineering* 56.3 (2017), págs. 031224-031224.

[12] Tobias Gruber et al. «Pixel-accurate depth evaluation in realistic driving scenarios». En: *2019 International Conference on 3D Vision (3DV)*. IEEE. 2019, págs. 95-105.

modela la transmisión atmosférica para separar las variables físicas implicadas en la observación.[13] De manera complementaria, se han propuesto enfoques que integran tanto la emisión del aire como la absorción atmosférica para estimar conjuntamente la profundidad y las propiedades del objeto, incorporando además estrategias orientadas a mejorar la robustez frente a la radiación incidente reflejada.[14] Por otra parte, existen arquitecturas basadas en aprendizaje profundo que descomponen las imágenes hiperespectrales en propiedades físicas como la temperatura, la emisividad y los factores de iluminación (textura), mediante funciones de pérdida fundamentadas en modelos de formación de imagen, aunque se omite la contribución del aire.[15] No obstante, los enfoques actuales aún presentan brechas importantes al no integrar de forma conjunta todos los componentes físicos relevantes y el potencial del aprendizaje profundo, lo que restringe su aplicabilidad y precisión en escenarios reales.

Por ello, en el presente trabajo se propone desarrollar y validar un algoritmo que combina modelos físicos de propagación de la luz a través de la atmósfera con técnicas de aprendizaje profundo, con el fin de estimar los mapas de profundidad de las escenas a partir de imágenes hiperespectrales adquiridas de manera pasiva en el LWIR. En este algoritmo, la imagen hiperespectral ingresa a la arquitectura de aprendizaje profundo, la cual aprende el mapeo a los parámetros físicos del modelo de formación de imagen en la etapa de entrenamiento mediante una función de costo basada en la física. La metodología propuesta incluye el modelamiento matemático del proceso de formación de imágenes hiperespectrales considerando los efectos de la emisión y la absorción atmosférica; posteriormente, la construcción de

---

[13] Nagase et al., ver n. 8.

[14] Unay Dorken Gallastegi et al. «Absorption-based, passive range imaging from hyperspectral thermal measurements». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[15] Fanglin Bao et al. «Heat-assisted detection and ranging». En: *Nature* 619.7971 (2023), págs. 743-748.

la arquitectura de aprendizaje profundo que integra dichas restricciones físicas; y finalmente, se evalúa el desempeño del enfoque propuesto utilizando un conjunto de datos sintéticos, así como una validación sobre datos experimentales, comparando los resultados frente a métodos de referencia del estado del arte donde se encuentra un rendimiento semejante que demuestra la viabilidad del método así como su complejidad y oportunidades de mejora para trabajo futuro.

# 1. OBJETIVOS

## Objetivo general

Desarrollar y validar un algoritmo que combine modelos físicos de propagación de la luz a través de la atmósfera con técnicas de aprendizaje profundo para estimar pasivamente la profundidad a partir de imágenes hiperespectrales en el infrarrojo de onda larga (LWIR).

## Objetivos específicos

1. Modelar matemáticamente el proceso de formación de imágenes hiperespectrales en el infrarrojo de onda larga (LWIR) considerando la propagación de la luz a través de la atmósfera.

2. Diseñar una arquitectura de aprendizaje profundo para la estimación pasiva de la profundidad a partir de imágenes hiperespectrales en el infrarrojo de onda larga (LWIR).

3. Implementar en Python la arquitectura de aprendizaje profundo diseñada, proponiendo una función de costo basada en el modelo de formación de imagen para guiar su entrenamiento mediante restricciones físicas.

4. Evaluar el desempeño del enfoque propuesto en la estimación pasiva de la profundidad, comparándolo con métodos del estado del arte mediante datos simulados.

# 2. MARCO DE REFERENCIA

## 2.1. Estimación pasiva de la profundidad

La estimación de la profundidad es el proceso de obtener información sobre la distancia de los puntos en una escena con respecto al sensor de la cámara. Inspirada en la percepción estereoscópica humana, esta tarea busca reconstruir la geometría tridimensional a partir de una o varias imágenes. Un resultado común es el mapa de profundidad, una imagen donde cada píxel representa la distancia a la cámara. Tal como se aprecia en la Figura 1, esta puede codificarse directamente en unidades métricas o mediante medidas relativas. Su precisión depende del método empleado y es frecuente encontrar regiones con datos faltantes debido a oclusiones, falta de textura o reflejos.[16] Las aplicaciones de la estimación de la profundidad incluyen la navegación autónoma, donde vehículos y robots detectan obstáculos y planifican rutas;[17,18] la inspección estructural y la topografía, al generar modelos 3D de infraestructuras;[19,20] el mapeo de interiores y exteriores,[21,22] y la reconstrucción

---

16 Szeliski, ver n. 3.

17 Hadsell et al., ver n. 2.

18 Geiger, Lenz y Urtasun, ver n. 5.

19 Jeffrey S Deems, Thomas H Painter y David C Finnegan. «Lidar measurement of snow depth: a review». En: *Journal of Glaciology* 59.215 (2013), págs. 467-479.

20 Yvette Y Lin et al. «ThermalNeRF: Thermal Radiance Fields». En: *2024 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2024, págs. 1-12.

21 Lin et al., ver n. 6.

22 Cesar Cadena et al. «Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age». En: *IEEE Transactions on Robotics* 32.6 (2016), págs. 1309-1332.

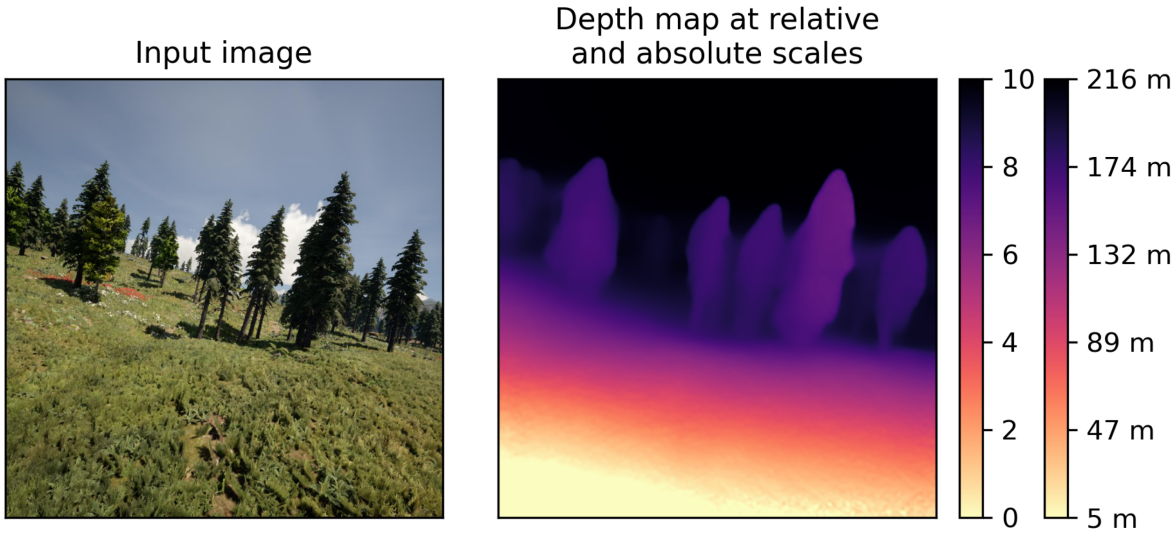


Figura 1. Ejemplo comparativo de profundidad relativa y absoluta: la imagen coloreada es la misma, pero los valores de cada píxel cambian de significado. La barra de la izquierda, sin unidades, indica profundidad relativa, la de la derecha, con unidades físicas, indica profundidad absoluta. Tomada de.[24]

tridimensional de objetos o escenas.[23]

A diferencia de los métodos activos, que emiten luz externa (por ejemplo, LiDAR o luz estructurada) y miden su interacción con los objetos para inferir la distancia,[25,26] la estimación pasiva se basa únicamente en la luz o la radiación ya presente en la escena, sin requerir iluminación externa. Entre ellos, la visión estéreo pasiva emplea múltiples cámaras calibradas para triangular puntos a partir de imágenes simultá-

---

[23] Ben Mildenhall et al. «Nerf: Representing scenes as neural radiance fields for view synthesis». En: *Communications of the ACM* 65.1 (2021), págs. 99-106.

[24] Chuanqi Zhang et al. «Monocular Absolute Depth Estimation from Motion for Small Unmanned Aerial Vehicles by Geometry-Based Scale Recovery». En: *Sensors (Basel, Switzerland)* 24.14 (2024), pág. 4541

[25] Miles Hansard et al. *Time-of-flight cameras: Principles, methods and applications*. Springer Science & Business Media, 2012.

[26] Chao Zuo et al. «Phase shifting algorithms for fringe projection profilometry: A review». En: *Optics and Lasers in Engineering* 109 (2018), págs. 23-59.

neas. Otros métodos explotan el desenfoque, estimando la profundidad según el grado de nitidez en la imagen. También existen técnicas que reconstruyen la forma a partir de las sombras o de patrones de textura, así como enfoques monoculares que infieren la geometría de la escena a partir de señales como la perspectiva, el tamaño relativo y la oclusión, con creciente apoyo de algoritmos de aprendizaje profundo.[27,28]

Sin embargo, todos estos métodos comparten limitaciones inherentes al depender de características visuales de la escena. Muchos requieren textura suficiente para establecer correspondencias fiables, lo cual dificulta el análisis en superficies homogéneas o bajo condiciones de poca luz.[29] La visión estéreo pasiva, en particular, pierde precisión a medida que aumenta la distancia debido a disparidades mínimas entre imágenes. Las oclusiones introducen zonas sin información confiable, mientras que superficies reflectantes o transparentes distorsionan la percepción de la geometría. En este contexto, la radiación térmica en el infrarrojo de onda larga (LWIR, del inglés, *LongWave InfraRed*) es particularmente prometedora debido a que la absorción varía con la distancia y la longitud de onda, ilustrado en la Figura 2, y esa dependencia se ha demostrado como una señal útil para inferir la profundidad.[30,31]

---

[27] Szeliski, ver n. 3.

[28] Lihe Yang et al. «Depth anything: Unleashing the power of large-scale unlabeled data». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 10371-10381.

[29] Szeliski, ver n. 3.

[30] Nagase et al., ver n. 8.

[31] Gallastegi et al., ver n. 14.

[32] Gallastegi et al., ver n. 14

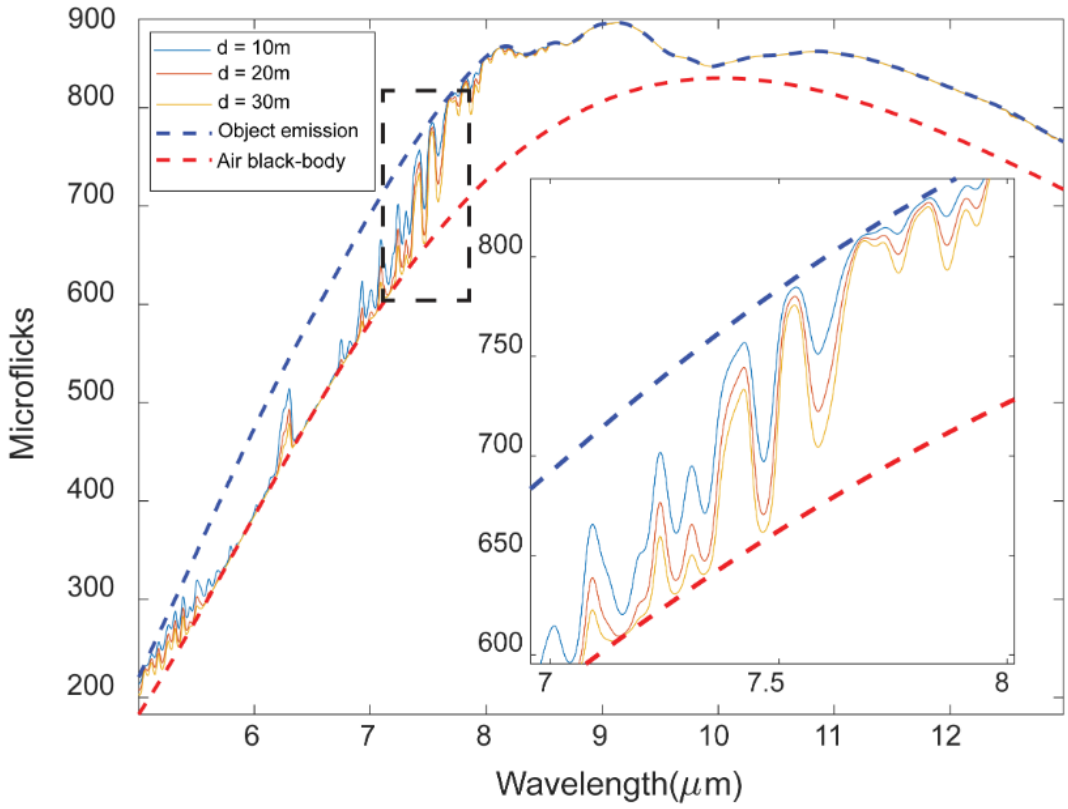


Figura 2. Simulación de medidas hiperespectrales (en microflicks) de la emisión de una roca a 300 K, registrada a 10 m (azul), 20 m (naranja) y 30 m (amarillo). Puede apreciarse cómo la absorción atmosférica, dependiente de la longitud de onda y de la distancia, atenúa la señal en bandas específicas. Tomada de.[32]

## 2.2. Radiación infrarroja

La termografía infrarroja se basa en las leyes de radiación de Kirchhoff y de Planck, que describen el espectro de la radiación térmica emitida por los objetos, y en la ley de Beer-Lambert, que modela la atenuación de esta radiación al atravesar un medio atenuante.[33] La ley de Planck establece que cualquier objeto con una temperatura superior a cero kelvin emite radiación. En este marco, se denomina *cuerpo negro* a

[33] Michael Vollmer y Klaus-Peter Möllmann. *Infrared thermal imaging: Fundamentals, research and applications*. John Wiley & Sons, 2018.

un emisor ideal que absorbe toda la radiación incidente. Planck describe la radiancia espectral emitida por un cuerpo negro como:

$$B(\lambda;T) = \frac{2 \times 10^{-4} hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}, \tag{1}$$

donde, $h = 6.626 \times 10^{-34}$ J·s es la constante de Planck, $c = 2.998 \times 10^8$ m/s la velocidad de la luz en el vacío, $\lambda$ la longitud de onda de la radiación, $T$ la temperatura absoluta del cuerpo y $k = 1.381 \times 10^{-23}$ J/K la constante de Boltzmann. Esta magnitud se expresa típicamente en microvatios por centímetro cuadrado, por estereorradián y por micrómetro de longitud de onda ($\mu W \cdot cm^{-2} \cdot sr^{-1} \cdot \mu m^{-1}$), también conocida como microflick.[34,35] Para cuerpos negros a 300 y 5777 K, la emisión máxima ocurre aproximadamente en 10 y 0.501 µm, respectivamente, como se muestra en la Figura 3. El primer caso se asemeja a la radiación ambiental, mientras que el segundo corresponde a la temperatura solar, cuyo pico está en el centro del espectro visible, por eso percibimos su luz como blanca.

Por otro lado, la ley de Kirchhoff sostiene que la radiación que absorbe un objeto es equivalente a la que emite, $\varepsilon = \alpha$, donde $\varepsilon$ representa la emisividad y $\alpha$ la absortividad, ambas medidas como fracciones de la radiación. Este principio se fundamenta en la conservación de la energía, ya que la radiación incidente sobre un objeto se divide entre lo que se refleja, lo que se transmite y lo que se absorbe, descrito mate-

---

[34] Gallastegi et al., ver n. 14.

[35] Bao et al., ver n. 15.

[36] Thomas G Pannuti. «Emission mechanisms: blackbody radiation, an introduction to radiative transfer, synchrotron radiation, thermal bremsstrahlung, and molecular rotational transitions». En: *The Physical Processes and Observing Techniques of Radio Astronomy: An Introduction*. Springer, 2020, págs. 69-114

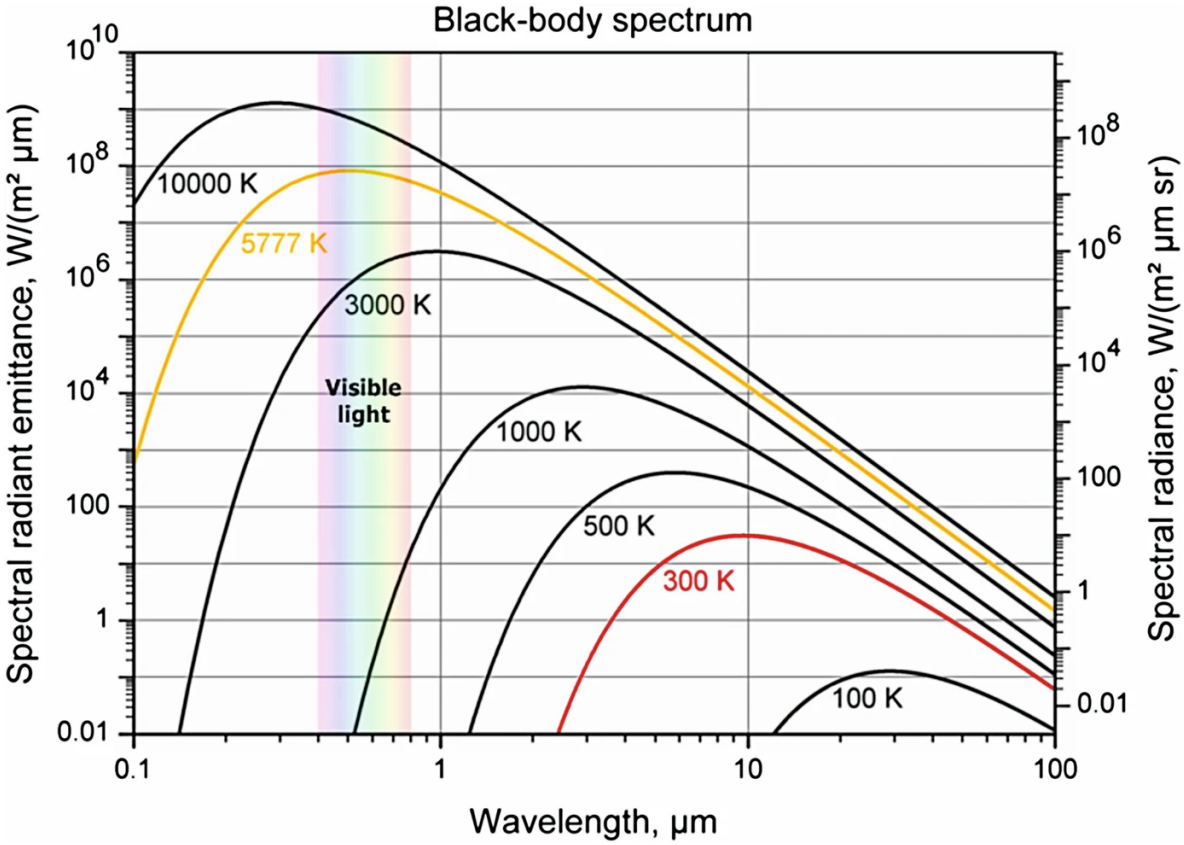


Figura 3. Espectro de radiación para cinco cuerpos negros con temperaturas de 100, 500, 1000, 3000 y 10000 K. También se incluyen dos perfiles adicionales para cuerpos negros con temperaturas de 300 y 5777 K (aproximando la emisión de cuerpo negro de la Tierra y el Sol, respectivamente). Tomada de.[36]

máticamente como $1 = \rho + \tau + \alpha$, siendo $\rho$ la reflectividad y $\tau$ la transmisividad.[37] La emisividad de un material no es constante y depende de varios factores como: el tipo de material, la estructura de la superficie, el ángulo de observación, la geometría, la longitud de onda y la temperatura.[38]

La ley de Beer-Lambert describe la atenuación de la radiación electromagnética al

---

[37] Vollmer y Möllmann, ver n. 33.

[38] Manolakis, Lockwood y Cooley, ver n. 9.

atravesar un medio. Esta ley se formula matemáticamente como

$$\tau(\lambda; d) = \frac{I_d(\lambda)}{I_0(\lambda)} = e^{-\gamma(\lambda)d}, \tag{2}$$

donde $\tau(\lambda; d)$ representa la transmitancia, $I_d(\lambda)$ la intensidad de la radiación tras recorrer una distancia $d$, $I_0(\lambda)$ la intensidad inicial, y $\gamma(\lambda)$ el coeficiente de atenuación total. Note la dependencia espectral de los términos.

En la óptica infrarroja, la ley de Beer-Lambert es esencial para comprender cómo la composición, la longitud de onda y el espesor del medio influyen en la atenuación de la radiación, y, junto con los fundamentos de las leyes de Planck y Kirchhoff, establece las bases para analizar la emisión, absorción y transmisión de la radiación.

### 2.3. Imágenes infrarrojas hiperespectrales

Las imágenes infrarrojas hiperespectrales captan información espectral detallada en múltiples bandas contiguas a lo largo de una porción del espectro electromagnético, típicamente en el infrarrojo de onda media (MWIR, del inglés *MidWave InfraRed*) y larga (LWIR, del inglés *LongWave InfraRed*).[39] A diferencia de las imágenes multiespectrales, las hiperespectrales comprenden cientos de bandas espectrales, lo que permite capturar un espectro, considerado continuo, para cada píxel de la imagen (véase la Figura 4). Esta riqueza de información espectral permite analizar la firma espectral representativa de los materiales y objetos presentes en la escena. Al analizar el espectro infrarrojo emitido o reflejado por un objeto en múltiples longitudes de onda se pueden inferir propiedades como la temperatura, la composición química y otras características físicas, lo que permite una identificación precisa en contraste

---

[39] Vollmer y Möllmann, ver n. 33.

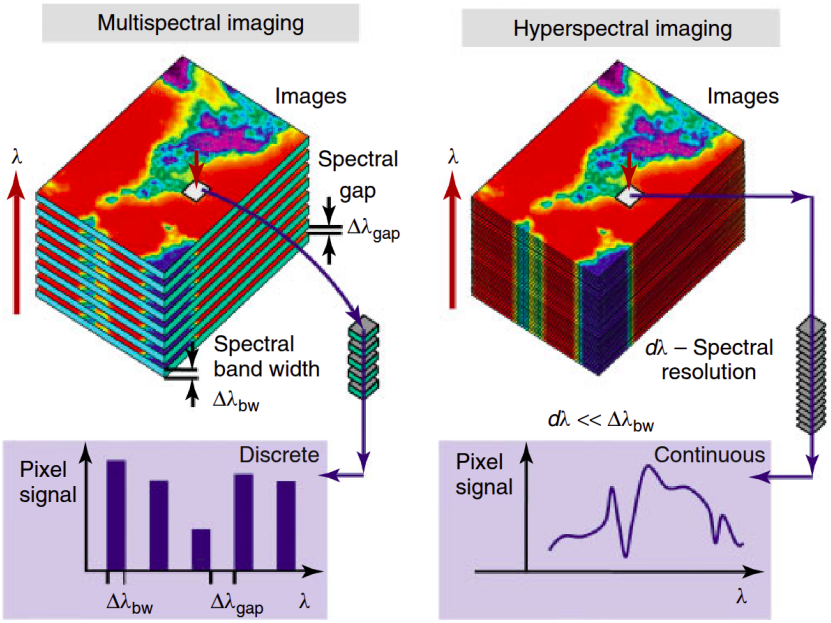


Figura 4. Comparación esquemática de las imágenes multiespectrales e hiperespectrales. La primera adquiere pocas bandas anchas y separadas, produciendo una señal espectral discreta. La segunda utiliza muchas bandas estrechas y contiguas, generando un espectro casi continuo por píxel. Tomada de.[41]

con las técnicas de banda ancha o multiespectrales.[40]

Esta riqueza espectral no solo permite identificar materiales, sino que también hace posible modelar de forma precisa el proceso de formación de imagen en el LWIR, lo cual es clave para tareas como la estimación pasiva de la profundidad.[42,43,44,45] En

---

[40] Dimitris Manolakis et al. «Longwave infrared hyperspectral imaging: Principles, progress, and challenges». En: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), págs. 72-100.

[41] Vollmer y Möllmann, ver n. 33

[42] Nagase et al., ver n. 8.

[43] Gallastegi et al., ver n. 14.

[44] Bao et al., ver n. 15.

[45] Takahiro Kushida et al. «Affine transform representation for reducing calibration cost on absorption-based LWIR depth sensing». En: *Scientific Reports* 14.1 (2024), pág. 26429.

este rango espectral, la radiancia espectral observada puede expresarse mediante la ecuación de transferencia radiativa,[46] que modela la combinación de la emisión del objeto y de la atmósfera a lo largo de múltiples longitudes de onda:[47]

$$L_{\mathrm{obs}}(\lambda) \;=\; \tau(\lambda; d)\, \varepsilon(\lambda)\, B(\lambda; T_{\mathrm{obj}}) \;+\; (1 - \tau(\lambda; d))\, B(\lambda; T_{\mathrm{air}})\,. \tag{3}$$

En esta expresión, $L_{\mathrm{obs}}(\lambda)$ es la radiancia espectral observada, $d$ es la distancia al objeto, $\varepsilon(\lambda)$ su emisividad espectral, y $T_{\mathrm{obj}}$ y $T_{\mathrm{air}}$ son las temperaturas del objeto y del aire, respectivamente. Asumiendo que $T_{\mathrm{air}}$ y el coeficiente de atenuación $\gamma(\lambda)$ de la Ecuación (2) son conocidos, este modelo puede invertirse para estimar simultáneamente la profundidad, la temperatura del objeto y su emisividad espectral,[48] formulando un problema de optimización de la siguiente forma:

$$\min_{d,\, T_{\mathrm{obj}},\, \varepsilon} \; \sum_{k=1}^{K} \big(\hat{L}_k(d, T_{\mathrm{obj}}, \varepsilon_k) - L_k\big)^2 \;+\; \beta \sum_{k=1}^{K-1} (\varepsilon_{k+1} - \varepsilon_k)^2\,, \tag{4}$$

donde las predicciones en las $K$ bandas espectrales, $\hat{L}_k$, se ajustan frente a las observadas $L_k$, y el parámetro $\beta$ impone una regularización que promueve la suavidad espectral en la emisividad.

Esta capacidad de caracterizar espectralmente los objetos ha sido fundamental para aplicaciones como la detección de sustancias químicas, el análisis de vegetación, la evaluación de daños en materiales y la caracterización de procesos indus-

---

46 Manolakis, Lockwood y Cooley, ver n. 9.

47 Gallastegi et al., ver n. 14.

48 Gallastegi et al., ver n. 14.

triales.[49,50,51,52] En este contexto, se han desarrollado algoritmos avanzados para tareas como la corrección atmosférica, la clasificación espectral, la extracción de información física y química y el análisis de series temporales, demostrando así el uso integral de las imágenes hiperespectrales en una amplia variedad de aplicaciones.[53,54,55]

## 2.4. Aprendizaje profundo guiado por la física

El aprendizaje profundo es un subcampo del aprendizaje automático basado en redes neuronales artificiales con múltiples capas de representación. Estas capas transforman los datos de entrada de manera progresiva, desde características básicas hasta representaciones más abstractas y útiles para tareas como la clasifi-

---

[49] Dimitris Manolakis, Steven Golowich y Robert S DiPietro. «Long-wave infrared hyperspectral remote sensing of chemical clouds: A focus on signal processing approaches». En: *IEEE Signal Processing Magazine* 31.4 (2014), págs. 120-141.

[50] Da-Wen Sun, Hongbin Pu y Jingxiao Yu. «Applications of hyperspectral imaging technology in the food industry». En: *Nature Reviews Electrical Engineering* 1.4 (2024), págs. 251-263.

[51] Prasad S Thenkabail, John G Lyon y Alfredo Huete. *Hyperspectral indices and image classifications for agriculture and vegetation*. CRC press, 2018.

[52] Chein-I Chang. *Hyperspectral imaging: Techniques for spectral detection and classification*. Vol. 1. Springer Science & Business Media, 2003.

[53] Dimitris Manolakis et al. «Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms». En: *IEEE Signal Processing Magazine* 31.1 (2013), págs. 24-33.

[54] José M Bioucas-Dias et al. «Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches». En: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5.2 (2012), págs. 354-379.

[55] Sicong Liu et al. «A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges». En: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), págs. 140-158.

cación, la segmentación y la estimación de la profundidad.[56] A diferencia de los métodos tradicionales de aprendizaje automático, que requieren ingeniería manual para extraer características relevantes de los datos crudos, el aprendizaje profundo puede descubrir automáticamente representaciones óptimas a partir de grandes conjuntos de datos mediante el entrenamiento de modelos usando algoritmos como el descenso de gradiente estocástico y la retropropagación del error.[57,58]

En particular, las redes neuronales convolucionales (CNN, por sus siglas en inglés, *Convolutional Neural Networks*) han demostrado ser especialmente efectivas para procesar datos espaciales, como imágenes, debido a sus operaciones de convolución y agrupamiento que capturan patrones espaciales jerárquicos.[59,60] Más recientemente, los *Transformers*,[61] y específicamente los transformadores de visión (ViT, del inglés, *Vision Transformer*), han emergido como arquitecturas poderosas en visión por computadora, prescindiendo de las convoluciones tradicionales, mediante mecanismos de auto-atención, ilustrado en la Figura 5. Los ViT dividen las imágenes en parches que se codifican como secuencias y se procesan a través de múltiples capas de atención multi-cabeza, lo que permite capturar dependencias espaciales complejas y mejorar significativamente el rendimiento en tareas como clasificación,

---

56 LeCun, Bengio e Hinton, ver n. 1.

57 LeCun, Bengio e Hinton, ver n. 1.

58 Yoshua Bengio, Yann Lecun y Geoffrey Hinton. «Deep learning for AI». En: *Communications of the ACM* 64.7 (2021), págs. 58-65.

59 LeCun, Bengio e Hinton, ver n. 1.

60 Alex Krizhevsky, Ilya Sutskever y Geoffrey E Hinton. «Imagenet classification with deep convolutional neural networks». En: *Advances in Neural Information Processing Systems* 25 (2012).

61 Ashish Vaswani et al. «Attention is all you need». En: *Advances in Neural Information Processing Systems* 30 (2017).

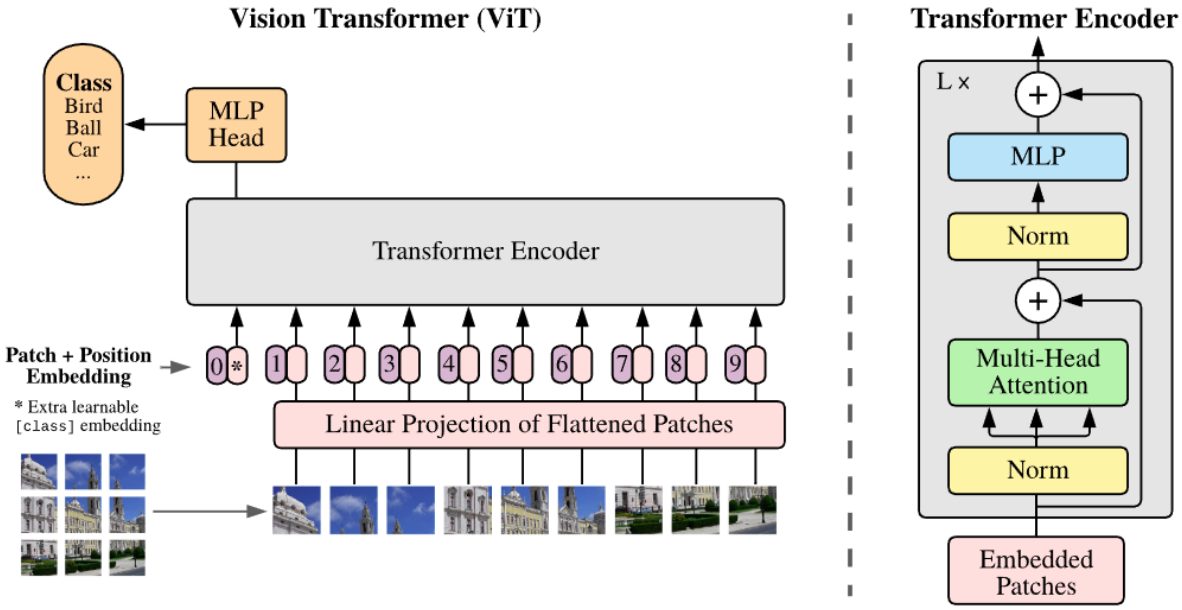


Figura 5. Flujo de procesamiento de un ViT: la imagen de entrada se fragmenta en parches, cada parche se proyecta linealmente y se le añade su codificación posicional; estos tokens (junto con el `[class]` token) pasan por $L$ bloques de *Transformer Encoder* con normalización, atención multi-cabeza y MLP (del inglés, *Multi-Layer Perceptron*) residuales. Tomada de.[65]

segmentación y estimación de la profundidad.[62,63,64]

La incorporación del conocimiento físico en el aprendizaje profundo ha surgido como un enfoque prometedor para mejorar la precisión y generalización de estos modelos. Estos enfoques utilizan funciones de pérdida informadas por modelos físicos del fenómeno subyacente, guiando al modelo hacia soluciones coherentes con las leyes

---

62 Yang et al., ver n. 28.

63 Alexey Dosovitskiy et al. «An image is worth 16x16 words: Transformers for image recognition at scale». En: *arXiv preprint arXiv:2010.11929* (2020).

64 Enze Xie et al. «SegFormer: Simple and efficient design for semantic segmentation with transformers». En: *Advances in Neural Information Processing Systems* 34 (2021), págs. 12077-12090.

65 Dosovitskiy et al., ver n. 63

físicas conocidas. Por ejemplo, en la estimación de la profundidad bajo el agua, se han propuesto arquitecturas específicas como PUDE (del inglés, *Physics-informed Underwater Depth Estimation*), que combinan modelos entrenados en aire y en ambientes submarinos, integrando restricciones físicas directamente en el proceso de aprendizaje. En la Figura 6 se ilustra este enfoque: durante el entrenamiento, imágenes subacuáticas se procesan mediante un modelo de referencia (DPT, por sus siglas en inglés, *Dense Prediction Transformer*)[66] y un modelo subacuático especializado (PUDE). A partir de las salidas del modelo de referencia, se extraen parámetros ópticos y se definen funciones de pérdida físicas y de transferencia que guían el aprendizaje, permitiendo así incorporar conocimiento físico sin depender de etiquetas de profundidad en las imágenes submarinas.[67]

Otros ejemplos notables incluyen redes neuronales que aceleran el cálculo del modelo de transferencia radiativa al aproximar modelos físicos,[69] la restauración y superresolución hiperespectral guiada por el modelo de formación de imagen,[70] la detección del metano con *Transformers* sensibles a sus bandas de absorción,[71] y

---

66 René Ranftl, Alexey Bochkovskiy y Vladlen Koltun. «Vision transformers for dense prediction». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 12179-12188.

67 Jinghe Yang, Mingming Gong y Ye Pu. «Physics-Informed Knowledge Transfer for Underwater Monocular Depth Estimation». En: *European Conference on Computer Vision*. Springer. 2024, págs. 449-465.

68 Yang, Gong y Pu, ver n. 67

69 Patrick G Stegmann et al. «A deep learning approach to fast radiative transfer». En: *Journal of Quantitative Spectroscopy and Radiative Transfer* 280 (2022), pág. 108088.

70 Yuchen Xiang et al. «Hyperspectral Image Restoration and Super-resolution with Physics-Aware Deep Learning for Biomedical Applications». En: *arXiv preprint arXiv:2503.02908* (2025).

71 Satish Kumar et al. «Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 17609-17618.

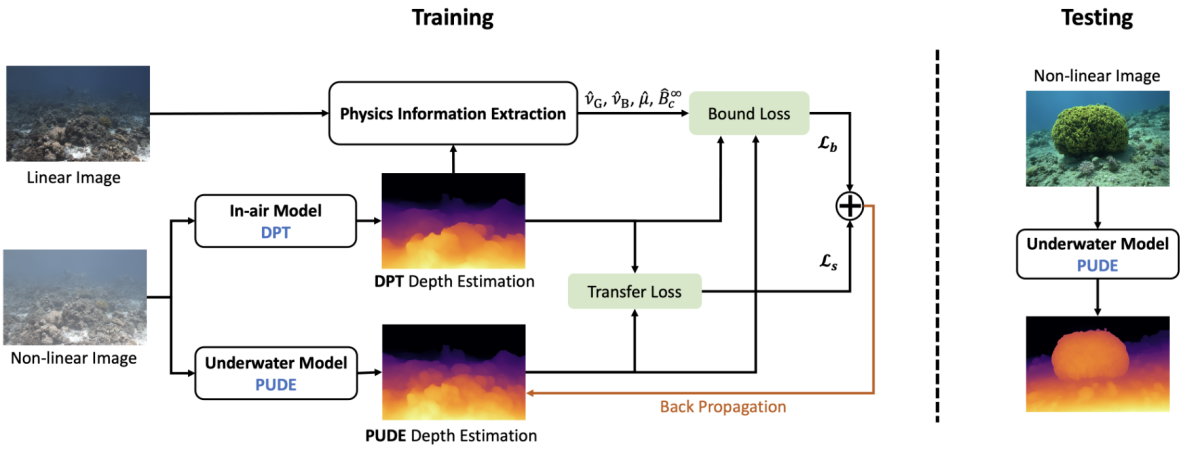


Figura 6. Esquema del enfoque propuesto en PUDE, donde las funciones de pérdida se derivan a partir de estimaciones generadas por un modelo de referencia entrenado en aire (DPT). Este enfoque permite entrenar el modelo sin necesidad de etiquetas de profundidad en imágenes submarinas. Tomada de.[68]

la estimación simultánea de varios contaminantes atmosféricos mediante redes informadas por leyes fisicoquímicas.[72] Estas metodologías destacan el potencial del aprendizaje profundo guiado por la física para abordar problemas complejos, especialmente en contextos donde la adquisición de datos etiquetados es limitada o costosa, como ocurre en escenarios hiperespectrales del LWIR. En este rango del espectro, la integración efectiva de conocimiento físico con modelos de aprendizaje profundo podría mejorar la precisión en tareas como la estimación pasiva de la profundidad.

[72] Binjie Chen et al. «An interpretable physics-informed deep learning model for estimating multiple air pollutants». En: *GIScience & Remote Sensing* 62.1 (2025), pág. 2482272.

# 3. MÉTODO PROPUESTO

El presente trabajo desarrolla un marco metodológico para la estimación pasiva de la profundidad a partir de imágenes hiperespectrales en el LWIR. La idea central consiste en aprovechar tanto el poder de los modelos de aprendizaje profundo, en particular las arquitecturas de tipo *Transformer* para predicción densa,[73] como las restricciones impuestas por el modelo físico de formación de imagen. Este enfoque híbrido permite no solo predecir mapas de profundidad consistentes, sino también garantizar que las estimaciones conserven coherencia con las propiedades físicas de la escena: la distancia, la temperatura y la emisividad de los materiales presentes.

El método propuesto se fundamenta en dos pilares complementarios. Por un lado, se desarrolla una arquitectura que integra un *Transformer encoder* pre-entrenado, adaptado para usar imágenes hiperespectrales, junto con tres *decoders* dedicados a la predicción densa de los mapas de profundidad, temperatura y emisividad. Por otro lado, estas predicciones se guían mediante una función de costo físico derivada del modelo de formación de imagen, la cual permite reconstruir la señal hiperespectral observada a partir de los atributos físicos estimados. De esta manera, la propuesta conecta el dominio del aprendizaje profundo con el modelado matemático, con la intención de que el proceso de estimación de la profundidad esté respaldado por la física del problema. La Figura 7 ilustra la idea general de la arquitectura propuesta, que se introduce a continuación.

---

[73] Ranftl, Bochkovskiy y Koltun, ver n. 66.

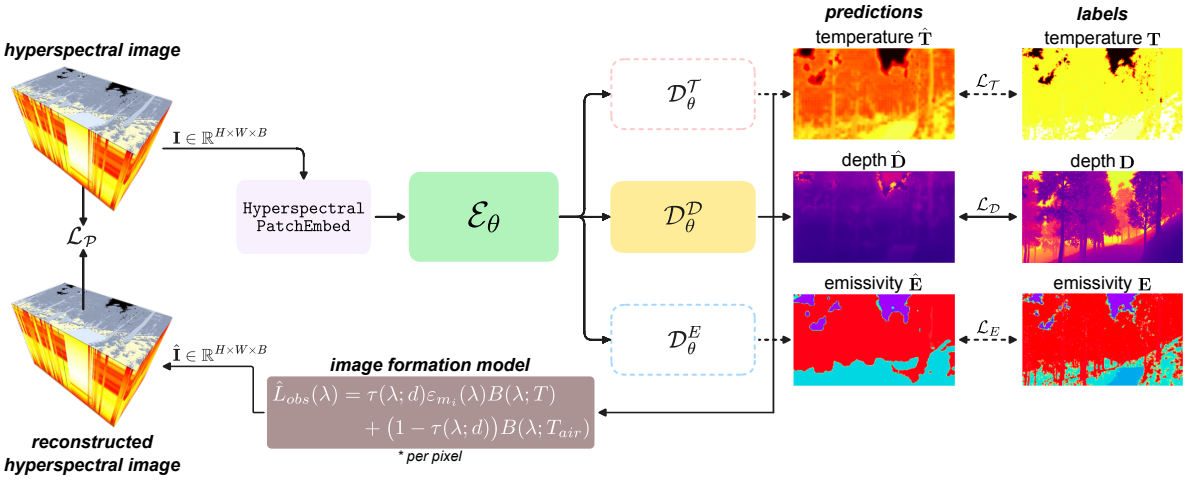


Figura 7. Esquema general de la arquitectura propuesta para la estimación de la profundidad a partir de imágenes hiperespectrales en el LWIR. Primero, la imagen hiperespectral $\mathbf{I} \in \mathbb{R}^{H\times W\times B}$ ingresa a la capa `HyperspectralPatchEmbed` donde es proyectada al grupo de *tokens* que ingresan al *Transformer encoder* $\mathcal{E}_\theta$. Luego, a partir de las representaciones latentes generadas, tres *decoders* especializados $\mathcal{D}_\theta^\mathcal{T}$, $\mathcal{D}_\theta^\mathcal{D}$ y $\mathcal{D}_\theta^E$ estiman, respectivamente, los mapas de temperatura, profundidad y emisividad de la escena. Finalmente, a partir de las predicciones se computan las pérdidas supervisadas $\mathcal{L}_\mathcal{T}$, $\mathcal{L}_\mathcal{D}$ y $\mathcal{L}_E$, y se reconstruye la radiancia espectral $\hat{\mathbf{I}} \in \mathbb{R}^{H\times W\times B}$ mediante el modelo de formación de imagen para calcular la pérdida física $\mathcal{L}_\mathcal{P}$, constituyendo un enfoque híbrido que conecta el aprendizaje profundo con la física del problema.

### 3.1. Modelo de formación de imagen

El punto de partida de nuestro método es el modelo de formación de imagen, el cual describe cómo la radiancia emitida por los objetos de la escena se transforma a lo largo del trayecto hacia el sensor. Este modelo permite establecer una relación explícita entre las propiedades físicas de interés (distancia, temperatura y emisividad) y la señal hiperespectral medida, proporcionando así la base para el planteamiento de la función de pérdida física que guía el aprendizaje de la arquitectura propuesta. Un cuerpo negro ideal a temperatura $T$ emite radiación electromagnética en la longitud de onda $\lambda$ de acuerdo con la ley de Planck, mostrada en la Ecuación (1). Sin embargo, los objetos reales no se comportan como cuerpos negros perfectos. La

emisividad $\varepsilon(\lambda)$ de un material se define como la relación entre la radiancia emitida por el objeto y la radiancia de un cuerpo negro a la misma temperatura. De esta forma, la emisión espectral de un material a temperatura $T$ se modela como

$$L_e(\lambda) = \varepsilon(\lambda)B(\lambda;T). \tag{5}$$

Durante su propagación hacia el sensor, parte de esta radiancia es absorbida por la atmósfera. La fracción de radiancia que alcanza el sensor depende de la transmitancia espectral $\tau(\lambda;d)$, que a su vez depende de la distancia $d$ recorrida (véase Ecuación (2)). Así, la contribución del objeto a la radiancia observada está dada por

$$L_{obj}(\lambda) = \tau(\lambda;d)\varepsilon(\lambda)B(\lambda;T). \tag{6}$$

Para materiales opacos, la reflectividad se relaciona con su emisividad mediante $\rho(\lambda) = 1 - \varepsilon(\lambda)$. La radiancia incidente sobre el objeto proviene de diversas fuentes ambientales. Asumiendo reflexión difusa de $N$ fuentes relevantes,[74] la radiancia reflejada se modela como

$$L_{ref}(\lambda) = \tau(\lambda;d)\left(1-\varepsilon(\lambda)\right)\sum_{i=1}^{N}\frac{\Omega_i}{\pi}L_{e,i}(\lambda), \tag{7}$$

donde $L_{e,i}(\lambda)$ corresponde a la radiancia espectral de la fuente $i$, y $\Omega_i$ a su ángulo sólido. En este caso, la transmitancia atmosférica entre el objeto y el sensor se representa nuevamente como $\tau(\lambda;d)$.

Por otra parte, la atmósfera también contribuye a la radiancia medida. Según la ley de Kirchhoff, bajo condiciones de equilibrio térmico la absortividad del aire es aproximadamente igual a su emisividad, de manera que $\varepsilon_{air}(\lambda) \approx 1 - \tau(\lambda;d)$. Aunque no

---

[74] Bao et al., ver n. 15.

todas las escenas naturales cumplen estrictamente este equilibrio, la aproximación es razonable cuando las variaciones de temperatura son moderadas. De esta forma, siguiendo la Ecuación (5), la radiancia emitida por el aire a una temperatura $T_{air}$ se expresa como

$$L_{air}(\lambda) = \big(1 - \tau(\lambda; d)\big) B(\lambda; T_{air}). \tag{8}$$

Finalmente, considerando conjuntamente la emisión del objeto, la reflexión ambiental y la contribución atmosférica, la radiancia espectral observada en el sensor se formula como

$$\begin{aligned} L_{obs}(\lambda) = & \underbrace{\tau(\lambda; d)\, \varepsilon(\lambda) B(\lambda; T)}_{\text{Emisión del objeto}} \\ & \underbrace{+ \tau(\lambda; d) \left(1 - \varepsilon(\lambda)\right) \sum_{i=1}^{N} \frac{\Omega_i}{\pi} L_{e,i}(\lambda)}_{\text{Radiancia reflejada}} \\ & + \underbrace{\big(1 - \tau(\lambda; d)\big) B(\lambda; T_{air})}_{\text{Emisión atmosférica}}. \end{aligned} \tag{9}$$

En el rango del LWIR, la radiancia solar reflejada suele tener una contribución despreciable, mientras que la emisión térmica de los objetos de la escena y de la atmósfera tiende a dominar.[75] Bajo este escenario, adoptamos la aproximación de conservar únicamente los términos dominantes de emisión del objeto y de la atmósfera, y omitir el término de radiancia reflejada. Esta hipótesis es especialmente válida en escenas naturales con objetos de alta emisividad, como la vegetación.

---

[75] Manolakis, Lockwood y Cooley, ver n. 9.

El modelo bajo esta asunción se simplifica a

$$L_{obs}(\lambda) = \underbrace{\tau(\lambda; d)\, \varepsilon(\lambda) B(\lambda; T)}_{\text{Emisión del objeto}} + \underbrace{\big(1 - \tau(\lambda; d)\big) B(\lambda; T_{air})}_{\text{Emisión atmosférica}}. \tag{10}$$

Esta ecuación constituye la base física de nuestro método, pues establece un vínculo directo entre las propiedades físicas de la escena y la señal medida por el sensor hiperespectral.

### 3.2. Adaptación del *Transformer encoder* pre-entrenado

En los últimos años, los modelos fundacionales basados en *Transformers* han demostrado un desempeño sobresaliente en tareas de visión por computadora a gran escala.[76,77] Estos modelos suelen ser pre-entrenados sobre bases de datos masivas de imágenes RGB (*Red, Green and Blue*, por sus siglas en inglés), y aprovechan una etapa inicial de proyección de las entradas al espacio de los *tokens* para poder procesarlas mediante un *encoder* de atención.

El mecanismo estándar de proyección se realiza a través de la capa `PatchEmbed`. Esta divide la imagen de entrada $\mathbf{J} \in \mathbb{R}^{H \times W \times C}$, (donde $H$ es la altura, $W$ la anchura y $C$ el número de canales, respectivamente) en parches no superpuestos de tamaño $p \times p$, generando un total de $N_p = HW/p^2$ parches. Cada parche se transforma mediante $D$ filtros convolucionales de tamaño $p \times p \times C$, donde $C$ es el número de canales de la entrada y $D$ es la dimensión de los *tokens* en el espacio latente del *Transformer*. Adicionalmente, se añade un *token* especial que se aprende y no está basado en la imagen de entrada, este *token* tiene el propósito de agregar la infor-

---

[76] Alexander Kirillov et al. «Segment anything». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, págs. 4015-4026.

[77] Lihe Yang et al. «Depth anything v2». En: *Advances in Neural Information Processing Systems* 37 (2024), págs. 21875-21911.

mación en una representación global de la imagen que se usa, generalmente, para clasificación.[78,79] El resultado de este proceso es un conjunto de *tokens* iniciales $t^0 = \{t_0^0, t_1^0, \ldots, t_{N_p}^0\}, \quad t_i^0 \in \mathbb{R}^D$, donde $t_0$ hace referencia al *token* especial añadido, que constituyen la representación de la imagen a la entrada del *encoder*.

Sin embargo, dado que los modelos fundacionales se entrenan originalmente sobre imágenes RGB, la capa `PatchEmbed` está configurada para operar con $C = 3$ canales. En nuestro caso, la entrada corresponde a imágenes hiperespectrales en el rango del LWIR, que poseen $B \gg 3$ bandas espectrales. Para hacer compatible la arquitectura pre-entrenada con este tipo de datos, fue necesario adaptar dicha capa.

Como se puede observar en la Figura 8, la estrategia seguida consistió en modificar los filtros iniciales de la capa `PatchEmbed` de la siguiente forma: (i) se replicaron los pesos de cada *kernel* pre-entrenado hasta cubrir el número de bandas $B$ de la imagen hiperespectral, y (ii) se reescalaron los valores resultantes multiplicando por el factor $3/B$, con el fin de mantener la magnitud de las activaciones comparable a la del caso RGB original. De este modo, cada parche de la imagen hiperespectral $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ se proyecta al espacio latente mediante $D$ filtros convolucionales de tamaño $p \times p \times B$. Este procedimiento, que denominamos `HyperspectralPatchEmbed`, permite reutilizar de manera efectiva el conocimiento del modelo fundacional pre-entrenado, al tiempo que ajusta la arquitectura a la dimensionalidad espectral del problema.

---

[78] Ranftl, Bochkovskiy y Koltun, ver n. 66.

[79] Dosovitskiy et al., ver n. 63.

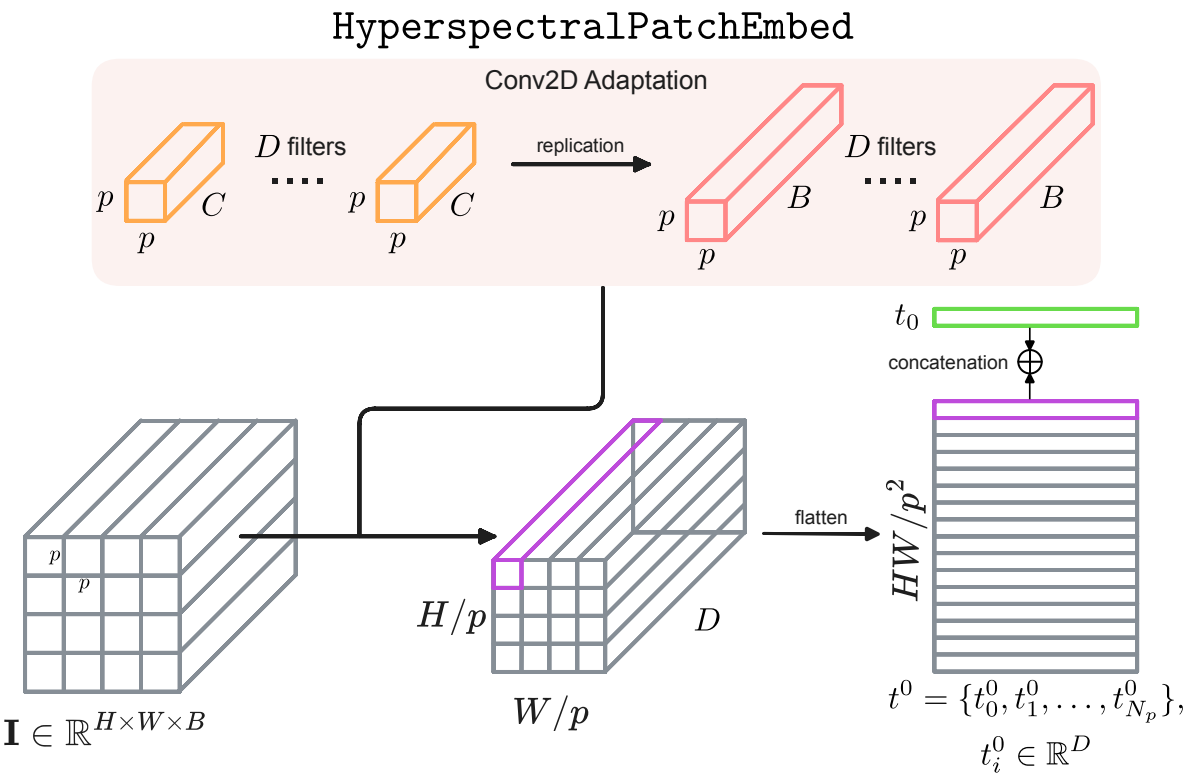


Figura 8. Diagrama de la capa `HyperspectralPatchEmbed`, empleada para adaptar la proyección inicial del *Transformer* pre-entrenado a imágenes hiperespectrales. El planteamiento consta de (i) replicar los pesos de cada *kernel* pre-entrenado hasta cubrir el número de bandas $B$, (ii) re-escalarlos mediante el factor $3/B$ para mantener la magnitud de las activaciones originales y, (iii) convolucionar la imagen $\mathbf{I} \in \mathbb{R}^{H\times W\times B}$ con los $D$ filtros para obtener el conjunto de *tokens* iniciales $t^0 = \{t_0^0, t_1^0, \ldots, t_{N_p}^0\}, \quad t_i^0 \in \mathbb{R}^D$, luego de añadir el *token* especial $t_0$, que forman la entrada del *encoder*.

### 3.3. Integración de los *decoders* de profundidad, temperatura y emisividad

La arquitectura propuesta integra un *Transformer encoder* pre-entrenado $\mathcal{E}_\theta$ que recibe como entrada la imagen hiperespectral $\mathbf{I} \in \mathbb{R}^{H\times W\times B}$ proyectada al conjunto de *tokens* iniciales $t^0$. A partir de las representaciones latentes generadas por este módulo, se conectan tres *decoders* especializados: $\mathcal{D}_\theta^{\mathcal{D}}$, $\mathcal{D}_\theta^{\mathcal{T}}$ y $\mathcal{D}_\theta^{E}$, diseñados para estimar, respectivamente, la profundidad, la temperatura y el mapa de emisividad de la escena.

En primer lugar, el *decoder* de profundidad $\mathcal{D}_\theta^{\mathcal{D}}$ produce un mapa

$$\hat{\mathbf{D}} = \mathcal{D}_\theta^{\mathcal{D}}\left(\mathcal{E}_\theta(\mathbf{I})\right), \qquad \hat{\mathbf{D}} \in \mathbb{R}^{H\times W}, \tag{11}$$

cuyos valores corresponden a la distancia en metros desde el sensor hasta cada píxel de la escena.

De manera análoga, el *decoder* de temperatura $\mathcal{D}_\theta^{\mathcal{T}}$ genera un mapa

$$\hat{\mathbf{T}} = \mathcal{D}_\theta^{\mathcal{T}}\left(\mathcal{E}_\theta(\mathbf{I})\right), \qquad \hat{\mathbf{T}} \in \mathbb{R}^{H\times W}, \tag{12}$$

expresado en grados Celsius. En ambos casos, las salidas se escalan considerando valores máximos de referencia obtenidos a partir del conjunto de datos, con el fin de mantener las predicciones dentro de rangos físicamente plausibles.

Por otra parte, tomando inspiración de,[80] el *decoder* de emisividad $\mathcal{D}_\theta^{E}$ produce un volumen

$$\hat{\mathbf{E}} = \mathcal{D}_\theta^{E}\left(\mathcal{E}_\theta(\mathbf{I})\right), \qquad \hat{\mathbf{E}} \in \mathbb{R}^{H\times W\times M}, \tag{13}$$

donde $M$ corresponde al número de materiales de la base de datos de emisividades. Cada canal representa la probabilidad de que el píxel pertenezca a una clase de material $m_i$. Posteriormente, se selecciona la clase más probable mediante el operador *argmax* sobre la dimensión de canales $M$, y se asigna la firma espectral de emisividad $\varepsilon_{m_i}(\lambda)$ correspondiente. Este procedimiento traduce la segmentación semántica de materiales en una representación espectralmente coherente de la emisividad.

Para guiar el entrenamiento de estos *decoders*, se definieron funciones de pérdida supervisadas específicas. En el caso de la profundidad y la temperatura, se empleó una versión comúnmente usada en la práctica de la *Scale-Invariant Logarithmic Loss*

---

[80] Bao et al., ver n. 15.

(*SiLogLoss*),[81] la cual resulta adecuada si queremos un balance entre la predicción en escala absoluta mientras conservamos la calidad cualitativa. Dada una predicción $\hat{y}_i$ y su valor de referencia $y_i$, se define como:

$$\text{SiLogLoss}(\hat{y}, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln \hat{y}_i - \ln y_i)^2 - \frac{\eta}{N^2}\left(\sum_{i=1}^{N}(\ln \hat{y}_i - \ln y_i)\right)^2}, \tag{14}$$

donde $N$ es el número de píxeles considerados y $\eta = 0.5$ (por defecto) es un hiperparámetro que balancea la consistencia global con la precisión local.

Aplicando esta métrica, las pérdidas para la profundidad y la temperatura se definen como:

$$\mathcal{L}_{\mathcal{D}} = \text{SiLogLoss}(\hat{\mathbf{D}}, \mathbf{D}), \qquad \mathcal{L}_{\mathcal{T}} = \text{SiLogLoss}(\hat{\mathbf{T}}, \mathbf{T}), \tag{15}$$

donde $\mathbf{D}$ y $\mathbf{T}$ representan los mapas de referencia de profundidad y temperatura, respectivamente.

En el caso de la emisividad, dado que se plantea como una tarea de segmentación semántica, se utilizó la función de pérdida *CrossEntropyLoss*, la cual mide la discrepancia entre los *logits* predichos $\hat{\mathbf{E}} \in \mathbb{R}^{H\times W\times M}$ y las etiquetas de clase discretas $\mathbf{E} \in \mathbb{N}^{H\times W}$. Para cada píxel $i$, esta pérdida se define como:

$$\mathcal{L}_E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbf{1}[E_i = m]\log\frac{\exp(\hat{E}_{i,m})}{\sum_{k=1}^{M}\exp(\hat{E}_{i,k})}, \tag{16}$$

donde $N = H \times W$ es el número de píxeles, $\hat{E}_{i,m}$ corresponde al *logit* predicho para la clase $m$ en el píxel $i$, y $\mathbf{1}[E_i = m]$ es un indicador que vale 1 si la clase verdadera en ese píxel es $m$ y 0 en caso contrario.

La combinación de estas pérdidas define la función de costo supervisada total, que

---

[81] David Eigen, Christian Puhrsch y Rob Fergus. «Depth map prediction from a single image using a multi-scale deep network». En: *Advances in Neural Information Processing Systems* 27 (2014).

integra la supervisión sobre los tres atributos físicos de manera conjunta:

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{\mathcal{D}} + \mathcal{L}_{\mathcal{T}} + \mathcal{L}_{E}. \tag{17}$$

Cabe resaltar que, tal como se ilustra en la Figura 7, los *decoders* $\mathcal{D}_{\theta}^{\mathcal{T}}$ y $\mathcal{D}_{\theta}^{E}$ aparecen representados con líneas punteadas, al igual que la conexión hacia sus respectivas pérdidas supervisadas. Esto obedece a que, dependiendo de la configuración de entrenamiento, el modelo puede operar en un modo reducido en el cual los mapas de temperatura, y emisividad, se proporcionan externamente para optimizar únicamente la estimación de la profundidad.

### 3.4. Planteamiento de la función de pérdida física

Si bien las pérdidas supervisadas introducidas permiten guiar el entrenamiento de los *decoders* hacia los mapas de referencia de profundidad, temperatura y emisividad, estas se limitan a aprovechar las etiquetas disponibles en el conjunto de datos. En escenarios reales, sin embargo, las etiquetas pueden ser escasas, incompletas o incluso inexistentes. Por esta razón, es fundamental incorporar restricciones físicas que actúen como una forma de regularización adicional, de modo que las predicciones del modelo no solo se ajusten a los datos de entrenamiento, sino que también respeten las leyes de la formación de imagen.

La idea básica consiste en utilizar las salidas estimadas por el modelo (profundidad $\hat{\mathbf{D}}$, temperatura $\hat{\mathbf{T}}$ y emisividad $\hat{\mathbf{E}}$) para reconstruir la imagen hiperespectral que debería observar el sensor, de acuerdo con la Ecuación (10). Posteriormente, esta radiancia reconstruida $\hat{\mathbf{I}} \in \mathbb{R}^{H\times W\times B}$ se compara con la imagen observada directamente $\mathbf{I} \in \mathbb{R}^{H\times W\times B}$. De este modo, se introduce un término de pérdida que penaliza discrepancias entre la física de la escena predicha y la señal espectral real. En este trabajo se empleó la norma $L_1$, su elección se debe a que ha sido ampliamente

utilizada en trabajos relacionados donde la tarea era reconstruir la señal original:[82,83]

$$\mathcal{L}_{\mathcal{P}} = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} \left| I^{(h,w,b)} - \hat{I}^{(h,w,b)} \right| \tag{18}$$

donde $B$ es el número de bandas espectrales y $H$, $W$ corresponden a las dimensiones espaciales de la imagen.

Finalmente, la función de pérdida híbrida total del modelo integra tanto las pérdidas supervisadas asociadas a las propiedades físicas como la pérdida física de reconstrucción espectral. Se define como:

$$\mathcal{L}_{\mathcal{H}} = \mathcal{L}_{\mathcal{S}} + \phi \mathcal{L}_{\mathcal{P}}, \tag{19}$$

donde el coeficiente $\phi$ permite ponderar la contribución relativa de la función de pérdida física. Esta formulación garantiza que el aprendizaje no solo esté guiado por la supervisión directa de etiquetas, sino también por la coherencia física entre las predicciones y la señal hiperespectral captada por el sensor.

---

[82] Bao et al., ver n. 15.

[83] Linus Scheibenreif, Michael Mommert y Damian Borth. «Masked vision transformers for hyperspectral image classification». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 2166-2176.

# 4. RESULTADOS

Los resultados de este trabajo primero cubren la identificación y adecuación de un conjunto de datos sintéticos de imágenes hiperespectrales en el LWIR a partir de los parámetros del modelo de formación de imagen. Segundo, una experimentación exhaustiva de los componentes de la arquitectura propuesta. Luego, la ejecución de pruebas comparativas frente a un método de referencia en el conjunto sintético para realizar un análisis cuantitativo y cualitativo. Y finalmente, con la indagación y reconocimiento de una base de datos experimental, se lleva a cabo la validación del enfoque propuesto ante métodos del estado del arte.

## 4.1. Base de datos

La base de datos *HADAR Database*[84] constituye un recurso fundamental en el estudio de imágenes hiperespectrales en el rango del LWIR. Esta colección está compuesta por 11 subconjuntos: diez de ellos son escenas sintéticas que representan condiciones de conducción autónoma (calles urbanas concurridas, carreteras, suburbios, zonas rurales, parques naturales, bosques, terrenos rocosos y desiertos), y un subconjunto adicional que corresponde a una escena experimental del mundo real.

Cada escena sintética fue generada en configuraciones estéreo (izquierda y derecha) con sensores montados en posiciones equivalentes a faros o en el techo de vehículos. Cada vista está formada por 5 fotogramas, y cada fotograma es una imagen hiperespectral de tamaño $1080 \times 1920 \times 54$ (altura, anchura y número de canales, respectivamente). En contraste, la escena experimental contiene 4 fotogramas de

[84] Bao et al., ver n. 15.

$260 \times 1500 \times 49$ (altura, anchura y número de canales, respectivamente).
Las imágenes hiperespectrales comparten un eje espectral definido en números de onda: $720 \sim 1250$ cm$^{-1}$ (o equivalentemente $8 \sim 13.8$ µm en longitud de onda) para los datos sintéticos, y $760 \sim 1240$ cm$^{-1}$ (o equivalentemente $8.1 \sim 13.2$ µm en longitud de onda) para los experimentales. Como puede verse en la Figura 9, la base de datos incluye además mapas de profundidad (en metros), mapas de temperatura superficial (en grados Celsius), mapas de emisividades (indices), y una base de datos de emisividades en el rango espectral de los datos experimentales, es decir, 49 bandas, para los 30 materiales que conforman las escenas. Cabe mencionar que todas las superficies son totalmente difusas.
A pesar de su riqueza, la base de datos *HADAR Database* no considera en sus renderizaciones el efecto de la absorción y emisión atmosférica. Dicho efecto es esencial para un modelado físico realista de la formación de la señal radiométrica, ya que el aire no es un medio perfectamente transparente en el rango LWIR. Ignorar esta contribución imposibilita utilizar directamente las imágenes hiperespectrales en el modelo de formación de imagen descrito en la Ecuación (10).

**4.1.1 Procedimiento de simulación** Para subsanar esta limitación, se diseñó un procedimiento que genera imágenes hiperespectrales sintéticas a partir de los mapas de emisividad, temperatura y profundidad, disponibles en la *HADAR Database*, el cual integra explícitamente los efectos de absorción y emisión atmosférica. El procedimiento se llevó a cabo en las siguientes etapas:

1. Conversión de unidades: Los mapas de temperatura superficial fueron convertidos de grados Celsius a kelvin, porque la función de cuerpo negro opera sobre kelvin.

2. Cálculo de la atenuación: La función de atenuación $\gamma(\lambda)$ se estimó usando el

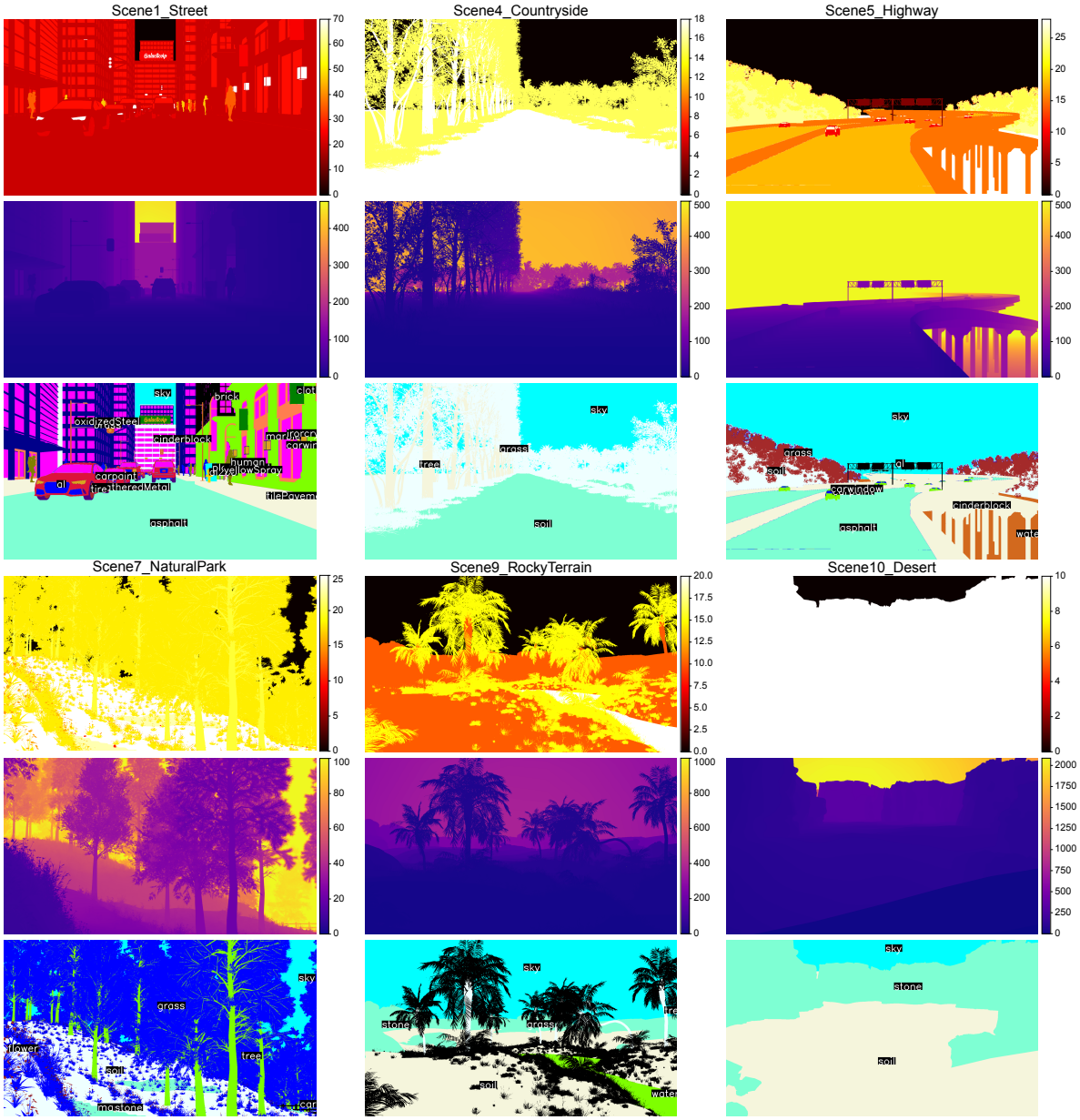


Figura 9. De arriba hacia abajo: mapas de temperatura, profundidad y emisividad para el primer fotograma de varias escenas sintéticas de la base de datos *HADAR Database*.

software *Spectral Calc*[85] a partir de condiciones atmosféricas estándar: atmós-

[85] Spectral Calc. [Online]. Available: `https://www.spectralcalc.com`.

fera a nivel del mar, temperatura ambiente de 15 °C, 4 % de humedad relativa y una columna de aire de 1 m. La atenuación fue tabulada con resolución espectral de 1 nm.

3. Ajuste de resoluciones espectrales: Como se puede observar en la Figura 10 (a) la emisividad de los materiales en la *HADAR Database* tiene una resolución aproximada de $\Delta\lambda = 123$ nm, lo que es incompatible con la resolución de la atenuación. Para resolver esta diferencia, se interpolaron y suavizaron las curvas espectrales de la atenuación original en la Figura 10 (b), empleando una ventana Gaussiana implementada mediante convolución. La atenuación ajustada resultante se puede visualizar en la Figura 10 (c).

4. Temperatura ambiente: Se asumió un valor fijo de $T_{air} = 293.15\,\text{K}$ (20 °C).

5. Modelo de formación de imagen: Se aplicó la Ecuación (10) para cada píxel, combinando la radiancia del objeto atenuada y la contribución del aire.

6. Aplicación de ruido: Para aproximar las condiciones de sensores en el LWIR reales, se añadió ruido aditivo Gaussiano con media cero y desviación estándar de un microflick, que corresponde al nivel de ruido común para sensores en este rango espectral.[86]

El resultado de este procedimiento es una nueva base de datos de imágenes hiperespectrales que mantiene la estructura espacial y de materiales de la *HADAR Database*, pero ahora incluye los efectos físicos atmosféricos necesarios para validar el modelo propuesto.

**4.1.2 Partición de la base de datos** Uno de los principales retos a la hora de realizar la validación experimental fue el número limitado de escenas disponibles en

[86] Bao et al., ver n. 15.

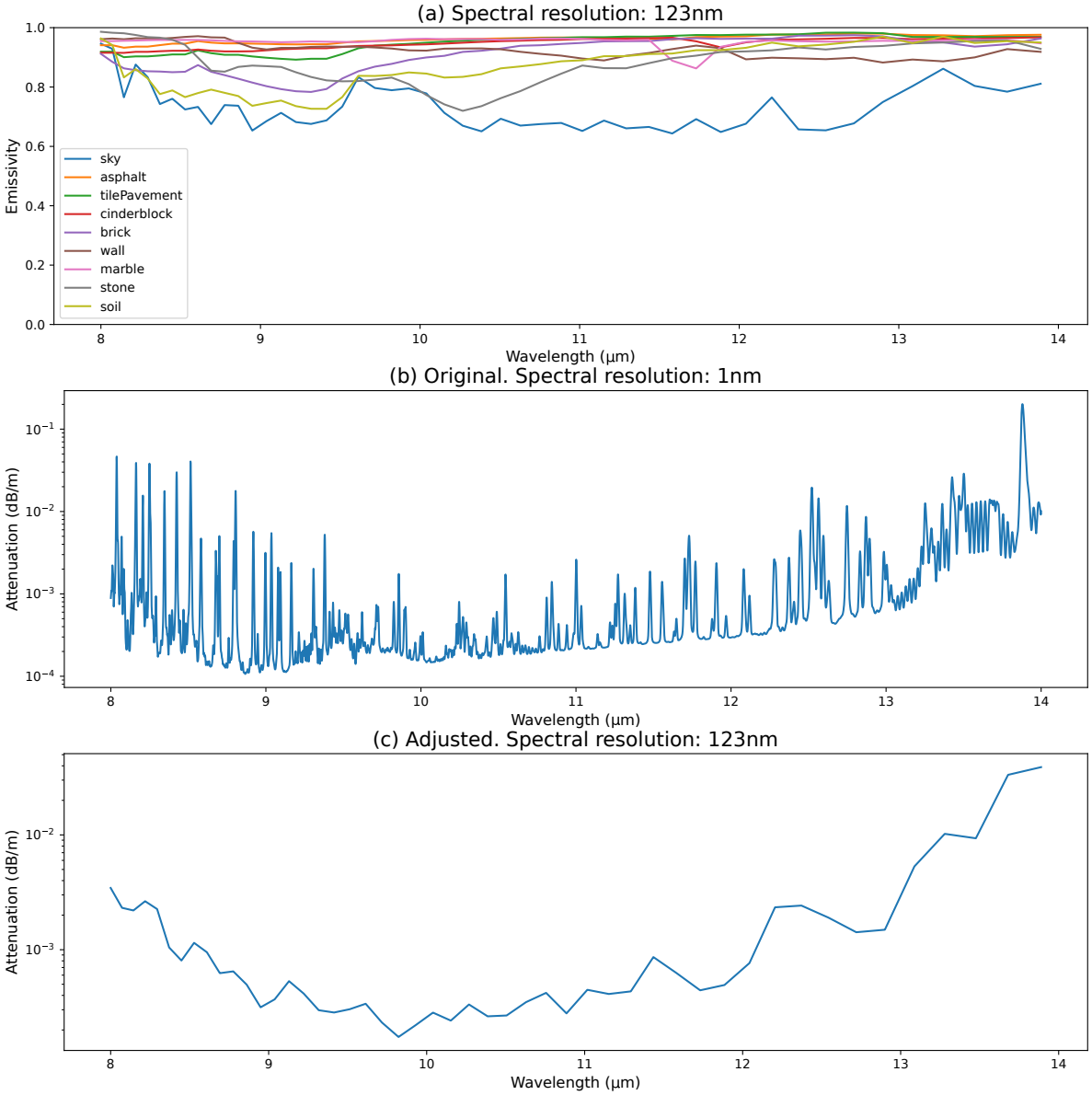


Figura 10. Relación entre emisividades y funciones de atenuación. (a) Ejemplos de perfiles de emisividad de materiales extraídos de la *HADAR Database*. (b) Función de atenuación atmosférica calculada con *Spectral Calc*, tabulada con resolución espectral de 1 nm. (c) Versión ajustada de la atenuación, interpolada y suavizada para coincidir con la resolución espectral (aprox. $\Delta\lambda = 123$ nm) de la base de datos de emisividades, garantizando la compatibilidad entre ambas fuentes de información espectral.

la base de datos. Contar con pocas muestras incrementa el riesgo de sobreajuste y dificulta obtener conclusiones generalizables acerca del desempeño de los métodos evaluados. Para mitigar este problema, se adoptó una estrategia de validación cruzada mediante *K-folds*, la cual permite aprovechar de manera más eficiente los datos disponibles y generar una estimación más robusta de la capacidad de generalización de los modelos.

En particular, se construyeron tres particiones distintas de los datos. En cada partición, seis escenas fueron asignadas para el entrenamiento y tres escenas fueron reservadas para la validación. De esta manera, cada modelo fue entrenado y evaluado tres veces, rotando el conjunto de validación en cada iteración. Las particiones definidas fueron las siguientes:

- **Fold 1:** validación con las escenas 1, 5 y 7.
- **Fold 2:** validación con las escenas 10, 2 y 9.
- **Fold 3:** validación con las escenas 3, 4 y 6.

La escena 8 (*Indoor*) se excluyó para limitar los experimentos a escenarios al aire libre. Este esquema de validación cruzada permitió reducir la dependencia de los resultados frente a una única división de entrenamiento/validación. Así, los experimentos no solo evalúan el desempeño medio de cada modelo en distintos subconjuntos de la base de datos, sino que también permiten identificar la variabilidad del error entre particiones, lo cual constituye un indicador importante de la estabilidad de los métodos comparados.

### 4.2. Métricas de evaluación

La evaluación del desempeño de algoritmos de estimación de la profundidad requiere el uso de métricas cuantitativas que permitan medir, de manera objetiva, la

calidad de los mapas generados. Estas métricas no solo facilitan la comparación entre distintos métodos, sino que también proporcionan información acerca de la magnitud y la naturaleza de los errores cometidos durante el proceso de estimación. A continuación, se describen las métricas empleadas en este trabajo.

**RMSE (del inglés, *Root Mean Squared Error*)**: Corresponde a la raíz cuadrada del promedio de los errores al cuadrado, definidos como la diferencia entre la profundidad real y la profundidad estimada en cada píxel del mapa. Esta métrica resulta especialmente sensible a errores grandes, por lo que constituye un indicador apropiado de la magnitud de las desviaciones en la estimación. El RMSE se expresa en metros y se define como:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i - \hat{d}_i\right)^2}, \tag{20}$$

donde $N$ representa el número de píxeles válidos considerados para la evaluación, $d_i$ es el valor de la profundidad verdadera y $\hat{d}_i$ el valor de la profundidad estimada en el píxel $i$. El conjunto de píxeles válidos está determinado por condiciones tales como la existencia de valores definidos tanto en el mapa real como en el estimado, y la pertenencia de dichos valores a un rango de profundidades físicamente plausible.

**AbsRel (del inglés, *Absolute Relative Error*)**: Mide la discrepancia relativa entre las profundidades verdaderas y estimadas, normalizando el error absoluto respecto a la profundidad real. De esta forma, la métrica refleja en qué proporción las estimaciones difieren del valor de referencia en términos porcentuales. Su expresión matemática es:

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - \hat{d}_i|}{d_i}. \tag{21}$$

En conjunto, estas métricas permiten obtener una visión integral del desempeño de

los algoritmos evaluados. Mientras que el RMSE enfatiza la magnitud de los errores, penalizando de manera más severa las grandes desviaciones, el AbsRel aporta una medida relativa. El uso combinado de ambas métricas ofrece, por tanto, una caracterización más completa de la calidad de los mapas de profundidad estimados.

## 4.3. Simulaciones

**4.3.1 Estudios de ablación** Con el fin de evaluar rigurosamente el rendimiento del método propuesto y validar la hipótesis de que la integración de un modelo físico del proceso de formación de imagen, considerando los efectos de emisión y absorción atmosférica, dentro de una arquitectura de aprendizaje profundo puede mejorar la estimación de la profundidad a partir de imágenes hiperespectrales en el LWIR, se llevó a cabo un estudio de ablación exhaustivo. Este análisis experimental permite identificar y cuantificar la contribución de cada componente de la arquitectura diseñada.

Como base para nuestros experimentos se eligió el modelo *Depth Anything V2*[87] que utiliza el esquema *encoder-decoder*. Este modelo ha sido recientemente propuesto como un modelo fundacional sólido para la estimación de profundidad monocular, destacándose por su capacidad de generalización y su facilidad para ser adaptado a tareas específicas mediante *fine-tuning*. Aunque disponible en las versiones *Small*, *Base* y *Large* con 24.8, 97.5 y 335.3 M de parámetros, respectivamente, elegimos la variante *Base*. Sus características hacen posible realizar un entrenamiento exhaustivo dentro de los recursos disponibles, manteniendo al mismo tiempo la competitividad del método.

La estructura modular de *Depth Anything V2* permite la incorporación de *decoders* adicionales, lo cual resulta especialmente apropiado para nuestro propósito: exten-

---

[87] Yang et al., ver n. 77.

der el modelo con salidas adicionales correspondientes a los mapas de temperatura y emisividad. Como *decoders* usamos las implementaciones de DPT (por sus siglas en inglés, *Dense Prediction Transformer*).[88] En este caso los tres *decoders* tienen la misma arquitectura. Sin embargo, la única diferencia es el número de canales en la última capa convolucional, los *decoders* de profundidad y temperatura predicen un solo canal que es modificado con la función de activación sigmoide y escalado por un valor determinado, usualmente el máximo absoluto de la magnitud escalar que se está estimando. En este contexto, el límite máximo de profundidad se estableció en 120 m, ligeramente por encima de la media registrada en la base de datos sintética, mientras que el límite máximo de temperatura se fijó en 70 °C, de acuerdo con el valor máximo observado en el conjunto. Por otro lado, el *decoder* de emisividad estima 30 canales (correspondientes al número de materiales en la base de datos), en el dominio de los *logits*.

Todos los experimentos fueron ejecutados en dos GPUs NVIDIA T4 a través de la plataforma Kaggle.[89] Se empleó un tamaño de lote (*batch size*) de 4 y un total de 30 épocas de entrenamiento. Durante el entrenamiento, cambiamos el tamaño de las imágenes para que el lado más pequeño tenga 420 píxeles, las recortamos aleatoriamente y las normalizamos usando la media y desviación estándar de la base de datos. En validación, los mapas de profundidad estimados fueron escalados y calculamos las métricas con la resolución original. Para optimizar el uso de memoria y acelerar el entrenamiento se implementó la técnica de precisión mixta (*mixed precision*) de 16 bits en los pesos del modelo.[90] El entrenamiento utilizó el optimizador

---

[88] Ranftl, Bochkovskiy y Koltun, ver n. 66.

[89] Kaggle. [Online]. Available: `https://www.kaggle.com`.

[90] Paulius Micikevicius et al. «Mixed precision training». En: *arXiv preprint arXiv:1710.03740* (2017).

AdamW,[91] con una tasa de aprendizaje diferenciada: $5 \times 10^{-6}$ para el *encoder*, a fin de preservar en lo posible los pesos pre-entrenados de alta calidad, y $5 \times 10^{-5}$ para los *decoders*. Tanto el *encoder* como el *decoder* de profundidad se inicializaron con los pesos de *Depth Anything V2* pre-entrenados en el conjunto de datos al aire libre Virtual KITTI 2,[92] mientras que los *decoders* de temperatura y emisividad, fueron inicializados de manera aleatoria.

Para garantizar una evaluación consistente, todos los modelos fueron validados utilizando el *fold* 1 de la base de datos sintética. La selección de este *fold* responde a que el conjunto de validación incluye tres escenarios contrastantes: una calle urbana, una carretera y un parque natural, lo que favorece la evaluación de la capacidad de generalización del modelo en entornos diversos.

El estudio de ablación se diseñó en torno a seis configuraciones principales, cada una de ellas destinada a aislar y analizar el efecto de los distintos componentes de la arquitectura:

- **Supervisado solo en distancia (D-S):** En este experimento se utilizó únicamente el *decoder* de profundidad, optimizando el mapa estimado de profundidad mediante la función de pérdida supervisada definida en la Ecuación (15).
- **Supervisado en distancia, emisividad y temperatura (DET-S):** En este caso, se añadieron los *decoders* correspondientes a temperatura y emisividad. Los tres mapas fueron optimizados empleando la función de pérdida supervisada descrita en la Ecuación (17).
- **Auto-supervisado en distancia, emisividad y temperatura (DET-P):** Para

---

[91] Ilya Loshchilov y Frank Hutter. «Decoupled weight decay regularization». En: *arXiv preprint arXiv:1711.05101* (2017).

[92] Yohann Cabon, Naila Murray y Martin Humenberger. «Virtual kitti 2». En: *arXiv preprint arXiv:2001.10773* (2020).

este experimento se incorporó la función de pérdida física propuesta en la Ecuación (18), con el objetivo de condicionar los mapas predichos a las restricciones del modelo físico de formación de imagen.

- **Auto-supervisado en distancia con soporte de emisividad y temperatura (D(ET)-P):** A diferencia del caso anterior, en este experimento se suministraron al modelo los mapas verdaderos de emisividad y temperatura. De este modo, la red debía optimizar únicamente la profundidad, pero manteniéndose coherente con el modelo físico mediante la pérdida física.

- **Híbrido en distancia, emisividad y temperatura (DET-H):** En el penúltimo experimento se combinó la pérdida supervisada y la pérdida física, de acuerdo con la Ecuación (19). Esta configuración busca alcanzar un equilibrio entre las ventajas del aprendizaje basado en datos y las restricciones impuestas por la física del problema.

- **Híbrido en distancia con soporte de emisividad y temperatura (D(ET)-H):** Finalmente, se exploró una variante híbrida en la que al modelo se le proveen los mapas reales de temperatura y emisividad, concentrándose únicamente en la optimización de la profundidad mediante la combinación de la pérdida supervisada en la Ecuación (15), y la pérdida física de la Ecuación (18).

Los resultados obtenidos para la estimación de la profundidad, a partir de estos experimentos de ablación se presentan en el Cuadro 1, donde se reportan los valores de RMSE y AbsRel alcanzados por cada configuración, en el conjunto de validación del *fold* 1.

| **Método** | **RMSE (m) ↓** | **AbsRel ( %) ↓** |
|---|---|---|
| D-S | **24.7148** | **0.5342** |
| DET-S | <u>26.5100</u> | 0.6355 |
| DET-P | 40.3083 | 2.4370 |
| D(ET)-P | 38.0838 | 0.6966 |
| DET-H ($\phi = 1$) | 27.2850 | 0.6155 |
| D(ET)-H ($\phi = 1$) | 27.1744 | <u>0.5754</u> |
| DET-H ($\phi = 15$) | 27.6196 | 0.7558 |
| D(ET)-H ($\phi = 15$) | 26.6100 | 0.5814 |

Cuadro 1. Resultados del estudio de ablación para la estimación de la profundidad. Se comparan seis configuraciones experimentales de la arquitectura propuesta: supervisado en distancia (D-S), supervisado en distancia, emisividad y temperatura (DET-S), auto-supervisado completo (DET-P), auto-supervisado con soporte de emisividad y temperatura (D(ET)-P), híbrido completo (DET-H) e híbrido con soporte (D(ET)-H). Se reportan los valores de RMSE y AbsRel obtenidos en el conjunto de validación del *fold* 1. El mejor resultado se muestra en negrita mientras que el segundo mejor se subraya.

Los resultados del estudio de ablación evidencian la complejidad del problema. Estimar simultáneamente los mapas de profundidad, emisividad y temperatura a partir de un conjunto de datos limitado resulta difícil, y en la práctica los mapas obtenidos tienden a perder texturas y detalles finos, incluso cuando se proporcionan como soporte los mapas verdaderos de emisividad y temperatura. Esta tendencia puede explicarse por la relativa uniformidad de los mapas en la base de datos, lo que reduce la información efectiva que aportan al modelo (véase Figura 9).

Como se puede apreciar en la Figura 11, en los experimentos con las configuraciones híbridas, específicamente D(ET)-H, encontramos que la magnitud de la pérdida supervisada de profundidad era aproximadamente quince veces mayor que la de la pérdida física. Al introducir un factor de compensación $\phi = 15$, el modelo híbrido

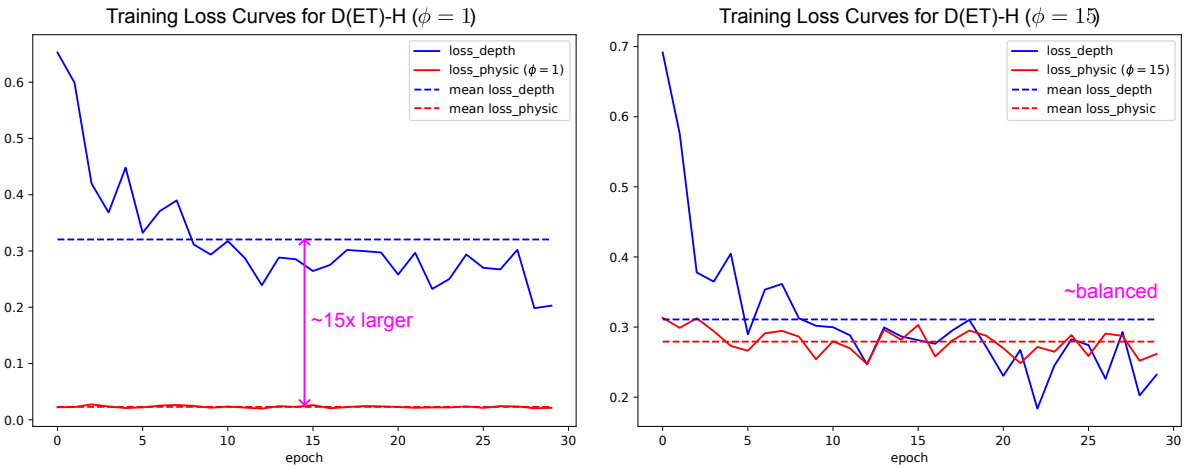


Figura 11. Curvas de pérdida en la etapa de entrenamiento para D(ET)-H con $\phi = 1$ y $\phi = 15$. La curva azul es la pérdida de profundidad $\mathcal{L}_{\mathcal{D}}$ y la roja la pérdida física $\mathcal{L}_{\mathcal{P}}$, a través de las épocas, mientras que las punteadas son la media.

logró un mejor balance entre ambas pérdidas, alcanzando un rendimiento competitivo. No obstante, el mejor desempeño en términos de RMSE y AbsRel se obtuvo al entrenar únicamente con la pérdida supervisada de profundidad (método (D-S) en el Cuadro 1). Esto sugiere que la arquitectura base, con pesos pre-entrenados, impone un sesgo fuerte que dificulta la incorporación de restricciones físicas de manera directa, lo que se evidencia en los resultados obtenidos con las configuraciones auto-supervisadas utilizando solamente la pérdida física. En consecuencia, los modelos híbridos no se alejan sustancialmente del rendimiento alcanzado por el modelo entrenado únicamente en profundidad, aunque aportan una mayor coherencia con el modelo físico subyacente.

Finalmente, a partir de estos hallazgos seleccionamos la configuración híbrida con soporte D(ET)-H ($\phi = 15$), como nuestra referencia principal al obtener el RMSE más bajo entre las configuraciones híbridas, y la utilizamos para comparar nuestro enfoque con los método de referencia en adelante.

**4.3.2 Resultados cuantitativos** La base de datos sintética desarrollada en este trabajo constituye, hasta donde tenemos conocimiento, el primer conjunto de datos de imágenes hiperespectrales en el rango del LWIR con atenuación y absorción atmosférica, para la estimación de propiedades físicas, como la distancia. Debido a ello, no existen métodos previos que integren de manera explícita el modelo físico de formación de imagen ni que exploten la información hiperespectral. En consecuencia, las comparaciones cuantitativas se realizaron contra enfoques de referencia basados en métodos de aprendizaje profundo de propósito general para la estimación de profundidad monocular.

Como punto de referencia se seleccionó nuevamente el modelo *Depth Anything V2* en su variante *Base*. Este modelo fue evaluado bajo tres esquemas de representación de la entrada, con el fin de explorar distintas formas de explotar la información espectral:

- **Pseudo-broadband (Sum):** Se obtuvo al sumar todos los canales espectrales de la imagen y normalizar el resultado con la estrategia *min-max normalization*. El mapa resultante, de un solo canal, se triplicó para generar una entrada pseudo-RGB compatible con el modelo de referencia.
- **Reducción mediante PCA:** En este caso, las imágenes hiperespectrales se estandarizaron canal por canal y posteriormente se aplicó análisis de componentes principales (PCA), reteniendo los primeros tres componentes principales a lo largo de los canales espectrales. El resultado se normalizó con *min-max normalization* para generar la entrada pseudo-RGB.
- **Entrada hiperespectral (HSI):** Se empleó la imagen hiperespectral en toda su dimensionalidad espectral como entrada, usando la estrategia propuesta en la Figura 8 para procesar directamente los 49 canales disponibles.

Para cada configuración se evaluaron dos escenarios. En primer lugar, se ejecu-

tó inferencia directa sobre los conjuntos de validación, empleando los pesos preentrenados del modelo. En segundo lugar, se realizó un proceso de *fine-tuning* sobre los conjuntos de entrenamiento de cada *fold*, con el fin de analizar el beneficio de una adaptación explícita a la base de datos propuesta. Adicionalmente, se incluyen los resultados de nuestra mejor configuración híbrida D(ET)-H ($\phi = 15$), descrita en la sección anterior, a fin de contrastar su desempeño con el del modelo de referencia.

Los resultados obtenidos luego de estas pruebas se incluyen en el Cuadro 2, donde se reportan los valores de la media y desviación estándar del RMSE y AbsRel obtenidos por cada modelo en los conjuntos de validación de cada *fold*.

| **Modelo** | **Entrada** | **RMSE (m)** $\downarrow$ | **AbsRel (%)** $\downarrow$ |
|---|---|---|---|
| Depth Anything V2 (Inference) | Sum | 25.2801 $\pm$ 6.9820 | **0.6353** $\pm$ **0.0955** |
| Depth Anything V2 (Inference) | PCA | 27.2339 $\pm$ 6.6338 | <u>0.6708 $\pm$ 0.1366</u> |
| Depth Anything V2 (Inference) | HSI | 25.2715 $\pm$ 6.6234 | 0.6778 $\pm$0.1677 |
| Depth Anything V2 (Fine-tuned) | Sum | <u>22.5860 $\pm$ 2.9759</u> | 0.8173 $\pm$ 0.4807 |
| Depth Anything V2 (Fine-tuned) | PCA | **21.9723** $\pm$ **2.9120** | 0.9690 $\pm$ 0.7237 |
| Depth Anything V2 (Fine-tuned) | HSI | 23.8792 $\pm$ 6.3643 | 0.8396 $\pm$ 0.4181 |
| D(ET)-H ($\phi = 15$) | HSI | 24.1020 $\pm$ 5.9249 | 0.8632 $\pm$ 0.4964 |

Cuadro 2. Comparación con enfoques de referencia para la estimación de la profundidad en la base de datos sintética. Se reportan los valores de la media y desviación estándar del RMSE y AbsRel obtenidos por cada modelo en los conjuntos de validación de cada *fold*. El mejor resultado se encuentra en negrita mientras que el segundo mejor se subraya.

Los resultados presentados en el Cuadro 2 permiten extraer varias conclusiones. En primer lugar, el modelo *Depth Anything V2* exhibe un rendimiento competitivo incluso sin ajuste, lo cual confirma su robustez como modelo fundacional. La estrategia de reducción mediante PCA demostró ser la más favorable tras el *fine-tuning*, alcanzan-

do el menor RMSE registrado (21.97 m), aunque a costa de un mayor error relativo (AbsRel). En contraste, las entradas pseudo-broadband (Sum) y HSI mostraron un mejor balance entre ambas métricas. Finalmente, la configuración híbrida D(ET)-H ($\phi = 15$), que integra restricciones físicas, obtuvo un desempeño comparable al de los enfoques de referencia finamente ajustados, lo que evidencia el potencial de combinar aprendizaje profundo con modelos físicos para estimar propiedades coherentes con el proceso de formación de imagen.

**4.3.3 Resultados cualitativos** El análisis cuantitativo, aunque indispensable, no logra capturar por completo la riqueza visual y las particularidades de las estimaciones de la profundidad. Por ello, complementamos dicho análisis con una evaluación cualitativa, en la cual se observan directamente las diferencias entre nuestro método y los enfoques de referencia basados en aprendizaje profundo.

En la Figura 12 se ilustran los resultados cualitativos obtenidos al emplear como entrada representaciones pseudo-RGB construidas a partir de la suma de canales (*Sum*) y mediante PCA. La primera columna corresponde a la imagen de entrada, seguida por las predicciones de los modelos de referencia, mientras que en la parte superior se incluye el *ground truth* para facilitar la comparación. El mapa de color se ajustó en el rango de cero a 120 m en todos los casos, pues este fue el valor usado para escalar las predicciones en el entrenamiento.

Se aprecia que, (i) los resultados de inferencia no logran discernir entre los objetos presentes en la escena y el fondo, resultando en un mapa de profundidad casi homogéneo en todos sus píxeles. (ii) Ajustar finamente los modelos en la base de datos sintética mejora notablemente los resultados y (iii) aunque los modelos de referencia logran capturar una estructura global coherente, presentan limitaciones en la representación de detalles finos, particularmente en elementos delgados como las ramas y los bordes de las hojas.

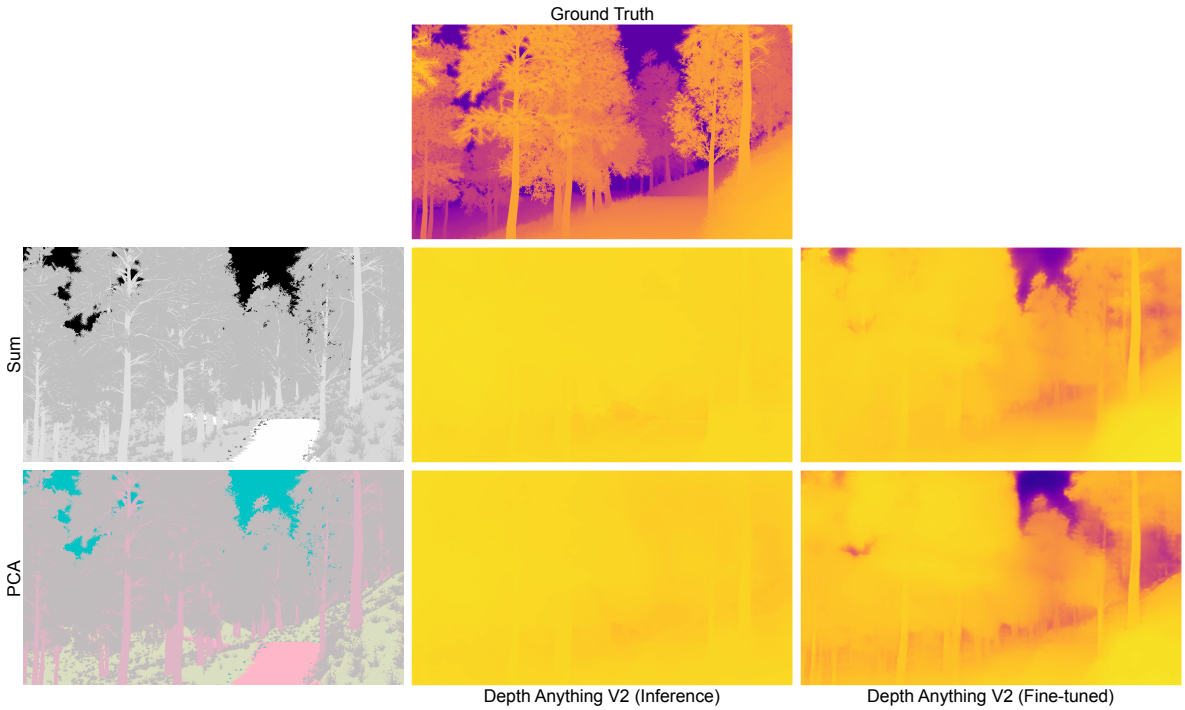


Figura 12. Comparación cualitativa de los resultados de estimación de la profundidad con entradas pseudo-RGB (*Sum* y PCA). De izquierda a derecha: imagen de entrada, predicciones en inferencia y resultados de ajuste fino. En la parte superior se muestra el *ground truth*.

Por otro lado, en la Figura 13 se muestran los resultados cualitativos obtenidos al emplear directamente la imagen hiperespectral (HSI) como entrada. Aquí se evidencia que el ajuste fino del modelo de referencia produce estimaciones regulares. Por ejemplo, se observan limitaciones en la detección de estructuras complejas, como las ramas de los árboles, y en la consistencia de la profundidad en regiones lejanas. Este último aspecto se debe en parte a la restricción impuesta en la base de datos, donde la distancia máxima confiable se estableció en 120 m.

En contraste, nuestra propuesta muestra un desempeño cualitativo superior. En azul, notamos que es competente para preservar detalles finos en la vegetación y en troncos en primer plano (primera fila), estructuras delgadas con formas regulares (segunda fila) y estructuras complejas con formas irregulares y varios elementos

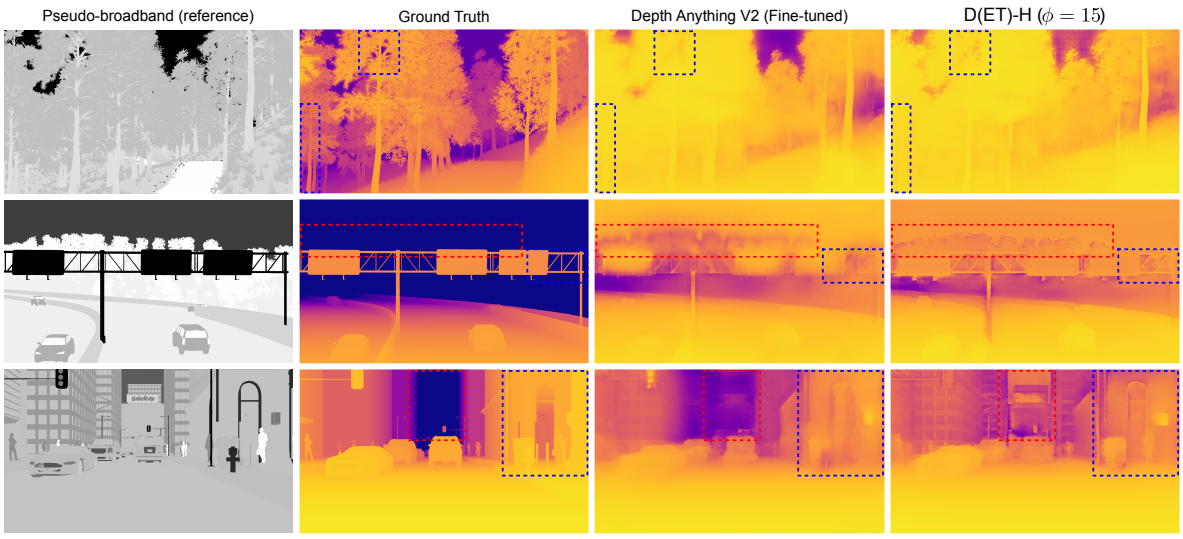


Figura 13. Comparación cualitativa de los resultados de estimación de la profundidad al emplear directamente la entrada hiperespectral (HSI). Por favor ampliar el documento para visualizar mejor los detalles. De izquierda a derecha: referencia visual pseudo-broadband, *ground truth*, predicción del método de referencia y estimación obtenida con nuestro método. En azul se comparan los detalles presentes y en rojo los objetos no presentes en el *ground truth*, respectivamente.

cerca (tercera fila). Adicionalmente, en rojo, hallamos que en escenas donde el mapa *ground truth* no presentaba valores de profundidad para ciertos objetos lejanos, nuestro método fue capaz de realizar una estimación confiable (segunda y tercera fila), lo que respalda la hipótesis de que la integración de información física en el modelo contribuye a mejorar la fidelidad de las estimaciones. No obstante, también se observa que los desafíos en distancias largas persisten, lo que abre un espacio para futuros trabajos en la extrapolación robusta de rangos extendidos.

En conjunto, estos resultados cualitativos complementan los hallazgos cuantitativos y refuerzan la conclusión de que nuestro método ofrece estimaciones de profundidad comparables y detalladas, especialmente en escenarios donde el modelo de referencia tiende a alucinar patrones o fallar en la preservación de estructuras.

**4.3.4 Resultados experimentales** Con el fin de evaluar la capacidad de generalización de nuestro modelo en escenarios del mundo real, realizamos una validación cualitativa y cuantitativa empleando imágenes provenientes de la base de datos *DARPA Invisible Headlights (IH) Dataset*.[93] Esta base de datos constituye un recurso único, pues recopila escenas *off-road* adquiridas con un conjunto diverso de sensores multiespectrales, hiperespectrales, polarimétricos y *broadband*, cubriendo un rango que se extiende desde el espectro visible hasta el LWIR. En particular, el sensor hiperespectral en el LWIR ofrece resoluciones de $1500 \times 260 \times 256$ o $1600 \times 260 \times 256$ (anchura, altura y número de canales, respectivamente).

Antes de emplear estas imágenes, fue necesario realizar un proceso de corrección espectral. Detectamos que existía un desalineamiento entre los centros de las longitudes de onda registradas en los metadatos y los correspondientes a la transmitancia atmosférica asumida. Para corregir este problema, aplicamos el algoritmo *AT²ES* con el objetivo de estimar la transmitancia en cada escena y compararla con la transmitancia teórica asumida.[94] Posteriormente, calculamos el coeficiente de correlación producto-momento de Pearson entre ambas transmitancias, utilizando ventanas estandarizadas. De esta forma, iteramos sobre diferentes candidatos de desplazamiento espectral y seleccionamos aquel que maximizaba el valor de correlación. Con dicho desplazamiento corregimos finalmente los metadatos de las imágenes. Además, cabe mencionar que el sensor hiperespectral empleado en la adquisición de las imágenes corresponde a un sistema de tipo *pushbroom*, que captura la escena mediante un barrido horizontal. Esta característica introduce arte-

---

[93] Florence Yellin et al. «Concurrent Band Selection and Traversability Estimation from Long-Wave Hyperspectral Imagery in Off-Road Settings». En: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, págs. 7483-7492.

[94] Sungho Kim, Jungsub Shin y Sunho Kim. «AT 2 ES: Simultaneous atmospheric transmittance-temperature-emissivity separation using online upper midwave infrared hyperspectral images». En: *Remote Sensing* 13.7 (2021), pág. 1249.

factos visibles en los datos recolectados, tales como líneas horizontales asociadas a respuestas no uniformes de los píxeles o a la presencia de elementos dañados en el detector. Para mitigar dichos efectos aplicamos un filtro de la mediana en la dirección horizontal.[95]

A pesar de la disponibilidad de nubes de puntos LiDAR (del inglés, *Light Detection And Ranging*) de alta resolución, no se encontraron etiquetas de profundidad. Por consiguiente, tuvimos que calibrar y registrar las nubes de puntos LiDAR con las imágenes hiperespectrales y de esta manera, poder realizar una comparación. El procedimiento primero consistió en comprender el modelo de la cámara de escaneo lineal giratorio,[96] el cual difiere del típico modelo de la cámara estenopeica para hacer la proyección de puntos 3D a mapas 2D.[97] Luego, utilizando el software *CloudCompare*[98] manualmente se anotaron varios puntos de correspondencia en las nubes de puntos LiDAR y las imágenes hiperespectrales, en lugares fácilmente distinguibles como las esquinas, bordes o vértices entre los objetos y el suelo. A partir de dichos puntos se plantea un problema de mínimos cuadrados no lineal para encontrar los intrínsecos y extrínsecos del modelo de la cámara utilizando como error la diferencia entre la reproyección de los puntos de correspondencia con los parámetros en la iteración actual y los manualmente anotados. Para finalizar, con los mejores parámetros se reproyectan las nubes de puntos LiDAR al espacio 2D obteniendo así un mapa de profundidad confiable para la evaluación. En la Figura 14 se pueden observar las imágenes pseudo-broadband, como referencia visual,

---

95 Bao et al., ver n. 15.

96 Fay Huang, Reinhard Klette y Karsten Scheibe. *Panoramic imaging: sensor-line cameras and laser range-finders*. John Wiley & Sons, 2008.

97 Szeliski, ver n. 3.

98 Daniel Girardeau-Montaut y CloudCompare Contributors. *CloudCompare (version 2.13.2)*. [Online]. Available: `https://www.cloudcompare.org`. 2024.

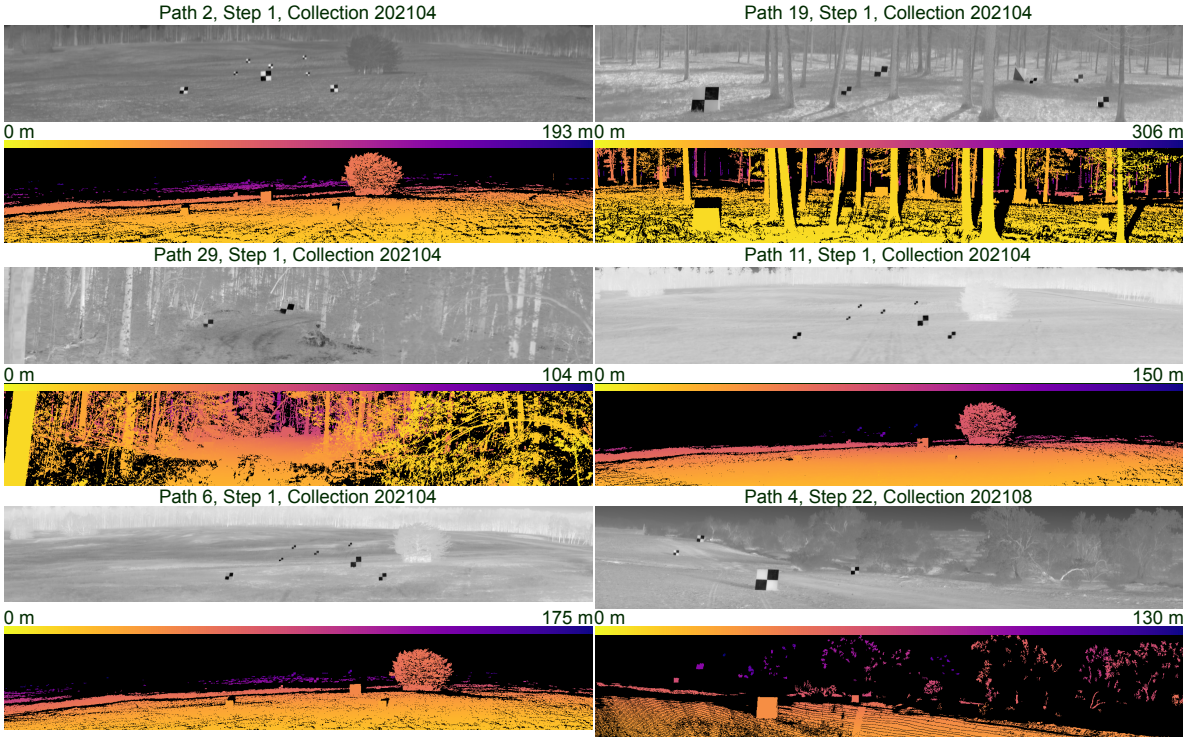


Figura 14. Imágenes pseudo-broadband y etiquetas de profundidad obtenidas para algunas escenas de la base de datos *IH Dataset*.

así como los mapas de profundidad elaborados para 6 escenas del *IH Dataset*. En estos, los píxeles negros representan puntos sin datos.

**Comparación cuantitativa** En la comparación cuantitativa, tomamos como métodos de referencia: (i) el promedio (también conocido como *ensemble*) de las predicciones usando los modelos *Depth Anything V2* finamente ajustados en los tres *folds* de la base de datos sintética, con entrada PCA. Esta selección se debe a que fue el método que obtuvo el mejor rendimiento en la comparación cuantitativa en dicho conjunto (véase Cuadro 2) y (ii) el propuesto en,[99] donde se optimizan mediante descenso del gradiente, por píxel, los tres parámetros de interés (temperatura, emisividad y distancia) del modelo de formación de imagen en la Ecuación (10) a través

[99] Gallastegi et al., ver n. 14.

del problema de optimización revisado en la Ecuación (4). Para esto replicamos el procedimiento especificado en el artículo. En nuestro caso, el cálculo de las derivadas se llevó a cabo utilizando diferenciación automática mediante la herramienta *autograd* de la librería de aprendizaje automático de código abierto *PyTorch*.[100] Para este método usamos las primeras 247 bandas de cada imagen, como se indicó por los autores del artículo. Cada píxel en la escena se optimizó durante 5000 iteraciones con una tasa de aprendizaje de $1 \times 10^{-2}$ para la temperatura y la emisividad, mientras que la distancia usó una tasa de aprendizaje más elevada de $1 \times 10^{5}$. El parámetro de regularización de la emisividad $\beta$ se estableció en $1 \times 10^{7}$. Ahora bien, el procedimiento con nuestro método D(ET)-H ($\phi = 15$) constó de muestrear cada imagen hiperespectral a 49 bandas usando una media móvil, con el fin de que puedan ser procesadas por los modelos entrenados en los tres *folds* de la base de datos sintética y así, computar su respectivo *ensemble*. Los resultados de la evaluación se exhiben en el Cuadro 3, donde se reportan la media del RMSE y AbsRel sobre las escenas etiquetadas. Para una comparación justa, las predicciones se evaluaron solo en los píxeles donde la profundidad en el *ground truth* LiDAR fuera menor que 120 m, pues esta fue la distancia máxima con la que se entrenaron los modelos.

| **Modelo** | **Entrada** | **RMSE (m)** $\downarrow$ | **AbsRel (%)** $\downarrow$ |
|---|---|---|---|
| Gradient Descent | HSI | 36.3802 | 0.9309 |
| Depth Anything V2 (Fine-tuned) | PCA | <u>23.7328</u> | <u>0.5636</u> |
| D(ET)-H ($\phi = 15$) | HSI | **21.2147** | **0.5031** |

Cuadro 3. Comparación con enfoques de referencia del estado del arte para la estimación de la profundidad en la base de datos *IH Dataset*. Se reporta la media del RMSE y AbsRel obtenidos por cada método en las escenas etiquetadas. El mejor resultado se encuentra en negrita mientras que el segundo mejor se subraya.

[100] Adam Paszke et al. «Automatic differentiation in PyTorch». En: *NeurIPS 2017 Workshop Autodiff* (2017).

Los resultados demuestran que el método propuesto D(ET)-H ($\phi = 15$), que integra información física en el entrenamiento de la arquitectura de aprendizaje profundo mediante una función de pérdida basada en el modelo de formación de imagen, contribuye a mejorar las métricas de las predicciones, presentándose como el más robusto ante escenarios de la vida real superando a los métodos del estado del arte.

**Comparaciones cualitativas** En la primera validación cualitativa nos comparamos con el método para la estimación de la profundidad a partir de imágenes hiperespectrales en el LWIR, presentado en el artículo HADAR (del inglés, *Heat Assisted Detection and Ranging*).[101] En este caso, utilizando la escena en el *Path* 27, *Step* 9 de la colección 202104 del *IH Dataset*, adaptada en el artículo HADAR en la dimensión espectral, de 256 a 49 canales. Los resultados cualitativos obtenidos se presentan en la Figura 15. Para efectos de visualización, se estableció la distancia máxima en el tercer cuantil del histograma de profundidades del mapa obtenido a partir del LiDAR. Primero, en la Figura 15 (a) se agrega una referencia visual pseudo-broadband sumando todos los canales, en la Figura 15 (b) se enseña el mapa de profundidad adquirido con LiDAR, en este, la correspondencia píxel a píxel es solo aproximada, pues el registro realizado en HADAR utilizó el modelo de la cámara estenopeica. En la Figura 15 (c), se utilizó la imagen *TeX* construida a partir de la salida de la red *TeXNet*, disponible en la base de datos *HADAR Database* como entrada al modelo *GCNDepth*,[102] replicando el procedimiento descrito en HADAR. La Figura 15 (d) es el *ensemble* por los modelos finamente ajustados en los tres *folds* de la base de datos sintética usando entrada PCA. Por último, en la Figura 15 (e), el *ensemble* de los modelos entrenados usando el enfoque propuesto.

En esta comparación visual puede apreciarse que nuestro modelo D(ET)-H ($\phi = 15$),

---

[101] Bao et al., ver n. 15.

[102] Armin Masoumian et al. «GCNDepth: Self-supervised monocular depth estimation based on graph convolutional network». En: *Neurocomputing* 517 (2023), págs. 81-92.

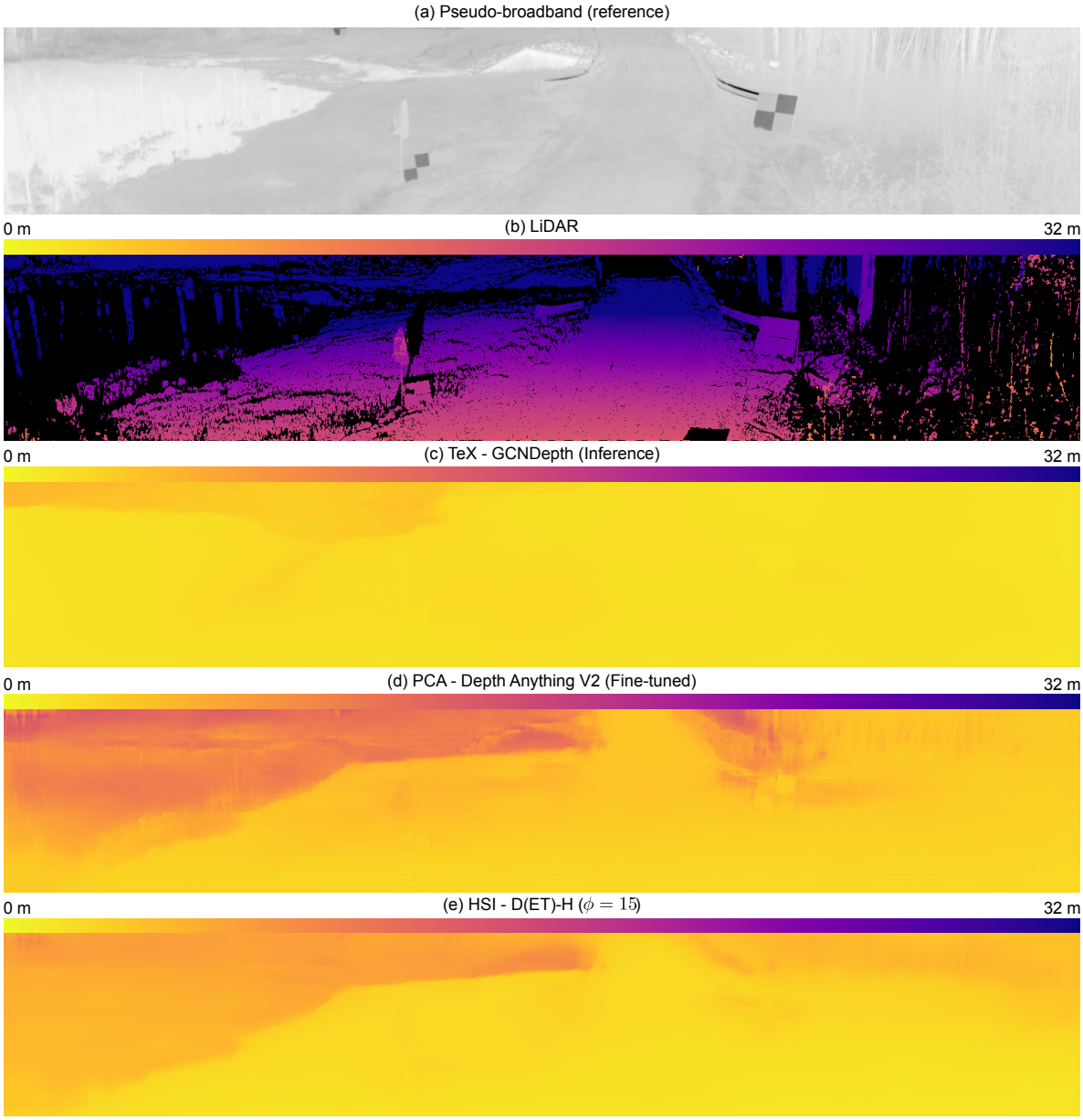


Figura 15. Comparación cualitativa de los resultados de estimación de la profundidad en una escena real de la base de datos *IH Dataset* frente al método empleado en HADAR. De arriba hacia abajo se ilustra: (a) referencia visual pseudo-broadband, (b) referencia LiDAR, (c) inferencia con los pesos pre-entrenados de *GCNDepth* con entrada *TeX*, método empleado en HADAR, (d) *ensemble* de *Depth Anything V2* finamente ajustado con entrada PCA y (e) *ensemble* de nuestro mejor modelo con entrada HSI.

proporciona una capacidad de generalización decente frente a escenas no vistas durante el entrenamiento. Sin embargo, el *ensemble* de *Depth Anything V2* finamente ajustado con entrada PCA ofrece una geometría más coherente. Aún así, ambos enfoques presentados en este trabajo muestran mayor robustez a la variabilidad propia de escenarios reales y complejos que el método presentado en HADAR.

Finalmente, en la segunda validación cualitativa realizamos una comparación frente al método del estado del arte,[103] donde proponen realizar la inversión del modelo mediante la Ecuación (4) usando el descenso del gradiente. Los resultados se enseñan en la Figura 16, utilizando la escena en el *Path 6, Step 1* de la colección 202104. Nuevamente, para apreciar mejor los detalles, se estableció la distancia máxima en el tercer cuantil del histograma de profundidades del mapa obtenido a partir del LiDAR. Para empezar, en la Figura 16 (a) se enseña la referencia visual pseudo-broadband. En la Figura 16 (b) se ilustra el mapa de profundidad obtenido del proceso de anotación y registro manual de este trabajo. Se puede observar que aunque el descenso del gradiente, en la Figura 16 (c), presenta mejor consistencia espacial, expone varias imprecisiones, por ejemplo, algunos parches en los patrones de calibración son estimados lejos, mientras que los troncos de los árboles son estimados cerca. El *ensemble* del ajuste fino de *Depth Anything V2* con entrada PCA en la Figura 16 (d), es el más robusto con los objetos en distancias largas, como los detalles en los troncos de los árboles en la parte superior derecha. En contraste, nuestra propuesta en la Figura 16 (e), ofrece mejores detalles en distancias cortas y medias, en particular, preserva la estructura de la mayoría de los patrones de calibración.

---

[103] Gallastegi et al., ver n. 14.

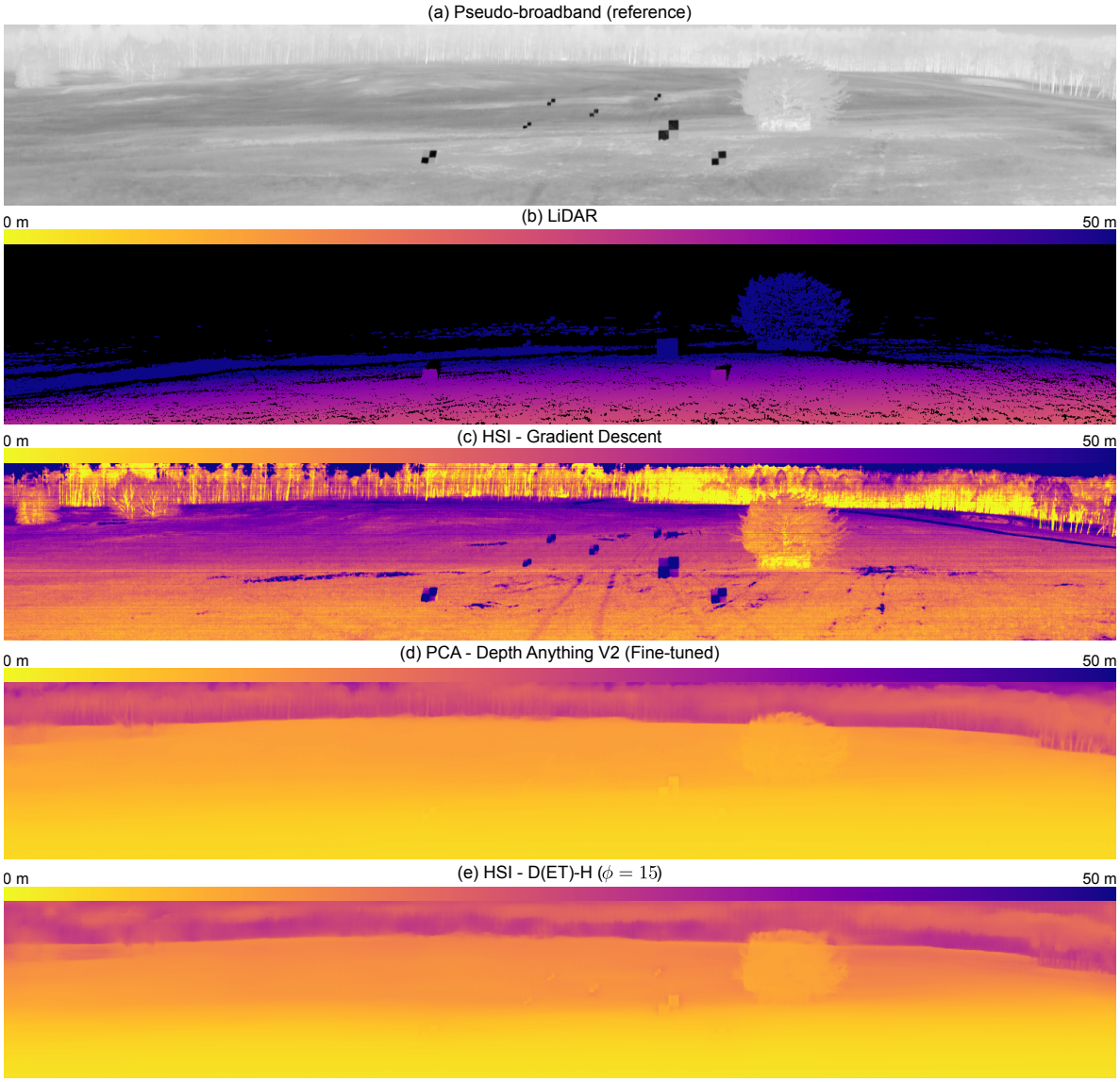


Figura 16. Comparación cualitativa de los resultados de estimación de la profundidad en una escena real de la base de datos *IH Dataset* frente al método empleado en *Absorption-Based, Passive Range Imaging From Hyperspectral Thermal Measurements*. De arriba hacia abajo se ilustra: (a) referencia visual pseudo-broadband, (b) referencia LiDAR, (c) predicción usando descenso del gradiente, método empleado en el artículo de referencia con entrada HSI, (d) *ensemble* de *Depth Anything V2* finamente ajustado con entrada PCA y (e) *ensemble* de nuestro mejor modelo con entrada HSI.

# 5. CONCLUSIONES

Este trabajo introdujo un enfoque híbrido para la estimación pasiva de la profundidad a partir de imágenes hiperespectrales en el rango infrarrojo de onda larga (LWIR), combinando la capacidad de las arquitecturas de aprendizaje profundo con el fundamento de un modelo físico de formación de imagen. La propuesta demostró que integrar un *Transformer encoder* pre-entrenado adaptado a imágenes hiperespectrales, con tres *decoders* para estimar conjuntamente los mapas de profundidad, temperatura y emisividad, guiado con una función de pérdida basada en el modelo de formación de imagen no solo obtiene resultados comparables a los métodos de referencia, sino que también procura su coherencia con las propiedades reales de la escena. La experimentación exhaustiva evidenció la complejidad del problema y se encontró que el balance entre las funciones de pérdida supervisada y física es fundamental. La evaluación cualitativa mostró que nuestra propuesta es capaz de preservar detalles finos y estructuras complejas en escenarios sintéticos. Por último, en la validación en el conjunto de datos *DARPA Invisible Headlights (IH) Dataset* se evidenció buena capacidad de generalización hacia escenarios no contemplados durante el entrenamiento, lo que resalta su potencial para aplicaciones en condiciones reales y variadas. En conjunto, los hallazgos sugieren que los modelos de aprendizaje profundo guiados por la física constituyen una vía prometedora para abordar problemas complejos de visión por computadora, como la estimación de la profundidad, y sientan bases sólidas para futuras investigaciones orientadas a la integración de conocimiento físico.

# 6. TRABAJO FUTURO

Los resultados del método propuesto abren varias líneas prometedoras para investigación futura. En primer lugar, el modelo de formación de imagen utilizado en este trabajo descansa sobre ciertas asunciones que, si bien facilitan su implementación, no siempre resultan válidas en entornos complejos. Entre ellas se encuentran la alta emisividad de los objetos o la homogeneidad de la atmósfera. Una extensión natural consiste en integrar de manera más completa aspectos introducidos en trabajos como HADAR (del inglés, *Heat Assisted Detection And Ranging*), donde se considera la reflectancia de los objetos, con el modelo empleado en este trabajo. Tal combinación permitiría obtener estimaciones conjuntas de distancia, temperatura y emisividad más precisas que las alcanzadas por cada enfoque por separado. Finalmente, no debe pasarse por alto que la información contenida en el espectro ofrece una oportunidad adicional: la estimación pasiva de parámetros atmosféricos. Explorar la posibilidad de inferir variables como la temperatura, la presión o la concentración de gases directamente a partir de la señal espectral representa un reto interesante. Resolver dicho problema no solo tendría un impacto en la caracterización ambiental, sino que también contribuiría a mejorar la estimación de la profundidad al eliminar la necesidad de asumir dichos parámetros como conocidos.

## BIBLIOGRAFÍA


Bao, Fanglin et al. «Heat-assisted detection and ranging». En: *Nature* 619.7971 (2023), págs. 743-748 (vid. págs. 14, 21, 24, 33, 38, 41, 42, 45, 61, 64).

Bengio, Yoshua, Yann Lecun y Geoffrey Hinton. «Deep learning for AI». En: *Communications of the ACM* 64.7 (2021), págs. 58-65 (vid. pág. 27).

Bioucas-Dias, José M et al. «Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches». En: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5.2 (2012), págs. 354-379 (vid. pág. 26).

Cabon, Yohann, Naila Murray y Martin Humenberger. «Virtual kitti 2». En: *arXiv preprint arXiv:2001.10773* (2020) (vid. pág. 51).

Cadena, Cesar et al. «Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age». En: *IEEE Transactions on Robotics* 32.6 (2016), págs. 1309-1332 (vid. pág. 17).

Cao, Zhexuan et al. «Aberration-robust monocular passive depth sensing using a meta-imaging camera». En: *Light: Science & Applications* 13.1 (2024), pág. 236 (vid. pág. 13).

Chang, Chein-I. *Hyperspectral imaging: Techniques for spectral detection and classification*. Vol. 1. Springer Science & Business Media, 2003 (vid. pág. 26).

Chen, Binjie et al. «An interpretable physics-informed deep learning model for estimating multiple air pollutants». En: *GIScience & Remote Sensing* 62.1 (2025), pág. 2482272 (vid. pág. 30).

Deems, Jeffrey S, Thomas H Painter y David C Finnegan. «Lidar measurement of snow depth: a review». En: *Journal of Glaciology* 59.215 (2013), págs. 467-479 (vid. pág. 17).

Dosovitskiy, Alexey et al. «An image is worth 16x16 words: Transformers for image recognition at scale». En: *arXiv preprint arXiv:2010.11929* (2020) (vid. págs. 28, 36).

Eigen, David, Christian Puhrsch y Rob Fergus. «Depth map prediction from a single image using a multi-scale deep network». En: *Advances in Neural Information Processing Systems* 27 (2014) (vid. pág. 39).

Gallastegi, Unay Dorken et al. «Absorption-based, passive range imaging from hyperspectral thermal measurements». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025) (vid. págs. 14, 19, 21, 24, 25, 62, 66).

Geiger, Andreas, Philip Lenz y Raquel Urtasun. «Are we ready for autonomous driving? the kitti vision benchmark suite». En: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, págs. 3354-3361 (vid. págs. 12, 17).

Girardeau-Montaut, Daniel y CloudCompare Contributors. *CloudCompare (version 2.13.2)*. [Online]. Available: `https://www.cloudcompare.org`. 2024 (vid. pág. 61).

Gruber, Tobias et al. «Pixel-accurate depth evaluation in realistic driving scenarios». En: *2019 International Conference on 3D Vision (3DV)*. IEEE. 2019, págs. 95-105 (vid. pág. 13).

Hadsell, Raia et al. «Learning long-range vision for autonomous off-road driving». En: *Journal of Field Robotics* 26.2 (2009), págs. 120-144 (vid. págs. 12, 17).

Hansard, Miles et al. *Time-of-flight cameras: Principles, methods and applications*. Springer Science & Business Media, 2012 (vid. pág. 18).

Huang, Fay, Reinhard Klette y Karsten Scheibe. *Panoramic imaging: sensor-line cameras and laser range-finders*. John Wiley & Sons, 2008 (vid. pág. 61).

Kaggle. [Online]. Available: `https://www.kaggle.com` (vid. pág. 50).

Kim, Sungho, Jungsub Shin y Sunho Kim. «AT 2 ES: Simultaneous atmospheric transmittance-temperature-emissivity separation using online upper midwave infrared hyperspectral images». En: *Remote Sensing* 13.7 (2021), pág. 1249 (vid. pág. 60).

Kirillov, Alexander et al. «Segment anything». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, págs. 4015-4026 (vid. pág. 35).

Krizhevsky, Alex, Ilya Sutskever y Geoffrey E Hinton. «Imagenet classification with deep convolutional neural networks». En: *Advances in Neural Information Processing Systems* 25 (2012) (vid. pág. 27).

Kumar, Satish et al. «Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 17609-17618 (vid. pág. 29).

Kushida, Takahiro et al. «Affine transform representation for reducing calibration cost on absorption-based LWIR depth sensing». En: *Scientific Reports* 14.1 (2024), pág. 26429 (vid. pág. 24).

LeCun, Yann, Yoshua Bengio y Geoffrey Hinton. «Deep learning». En: *Nature* 521.7553 (2015), págs. 436-444 (vid. págs. 12, 27).

Lin, Haotong et al. «Prompting Depth Anything for 4K Resolution Accurate Metric Depth Estimation». En: *arXiv preprint arXiv:2412.14015* (2024) (vid. págs. 12, 17).

Lin, Yvette Y et al. «ThermalNeRF: Thermal Radiance Fields». En: *2024 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2024, págs. 1-12 (vid. pág. 17).

Liu, Sicong et al. «A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges». En: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), págs. 140-158 (vid. pág. 26).

Loshchilov, Ilya y Frank Hutter. «Decoupled weight decay regularization». En: *arXiv preprint arXiv:1711.05101* (2017) (vid. pág. 51).

Manolakis, Dimitris, Steven Golowich y Robert S DiPietro. «Long-wave infrared hyperspectral remote sensing of chemical clouds: A focus on signal processing approaches». En: *IEEE Signal Processing Magazine* 31.4 (2014), págs. 120-141 (vid. pág. 26).

Manolakis, Dimitris et al. «Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms». En: *IEEE Signal Processing Magazine* 31.1 (2013), págs. 24-33 (vid. pág. 26).

Manolakis, Dimitris et al. «Longwave infrared hyperspectral imaging: Principles, progress, and challenges». En: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), págs. 72-100 (vid. pág. 24).

Manolakis, Dimitris G, Ronald B Lockwood y Thomas W Cooley. *Hyperspectral imaging remote sensing: Physics, sensors, and algorithms*. Cambridge University Press, 2016 (vid. págs. 13, 22, 25, 34).

Masoumian, Armin et al. «GCNDepth: Self-supervised monocular depth estimation based on graph convolutional network». En: *Neurocomputing* 517 (2023), págs. 81-92 (vid. pág. 64).

Micikevicius, Paulius et al. «Mixed precision training». En: *arXiv preprint arXiv:1710.03740* (2017) (vid. pág. 50).

Mildenhall, Ben et al. «Nerf: Representing scenes as neural radiance fields for view synthesis». En: *Communications of the ACM* 65.1 (2021), págs. 99-106 (vid. pág. 18).

Nagase, Yasuto et al. «Shape from thermal radiation: Passive ranging using multispectral lwir measurements». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 12661-12671 (vid. págs. 13, 14, 19, 24).

Pannuti, Thomas G. «Emission mechanisms: blackbody radiation, an introduction to radiative transfer, synchrotron radiation, thermal bremsstrahlung, and molecular rotational transitions». En: *The Physical Processes and Observing Techniques of Radio Astronomy: An Introduction*. Springer, 2020, págs. 69-114 (vid. pág. 21).

Paszke, Adam et al. «Automatic differentiation in PyTorch». En: *NeurIPS 2017 Workshop Autodiff* (2017) (vid. pág. 63).

Ranftl, René, Alexey Bochkovskiy y Vladlen Koltun. «Vision transformers for dense prediction». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 12179-12188 (vid. págs. 29, 31, 36, 50).

Saxena, Ashutosh, Sung H Chung y Andrew Y Ng. «3-d depth reconstruction from a single still image». En: *International Journal of Computer Vision* 76 (2008), págs. 53-69 (vid. pág. 12).

Scheibenreif, Linus, Michael Mommert y Damian Borth. «Masked vision transformers for hyperspectral image classification». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 2166-2176 (vid. pág. 41).

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. 12).

Spectral Calc. [Online]. Available: `https://www.spectralcalc.com` (vid. pág. 44).

Stegmann, Patrick G et al. «A deep learning approach to fast radiative transfer». En: *Journal of Quantitative Spectroscopy and Radiative Transfer* 280 (2022), pág. 108088 (vid. pág. 29).

Sun, Da-Wen, Hongbin Pu y Jingxiao Yu. «Applications of hyperspectral imaging technology in the food industry». En: *Nature Reviews Electrical Engineering* 1.4 (2024), págs. 251-263 (vid. pág. 26).

Szeliski, Richard. *Computer vision: Algorithms and applications*. Springer Nature, 2022 (vid. págs. 12, 17, 19, 61).

Thenkabail, Prasad S, John G Lyon y Alfredo Huete. *Hyperspectral indices and image classifications for agriculture and vegetation*. CRC press, 2018 (vid. pág. 26).

Vaswani, Ashish et al. «Attention is all you need». En: *Advances in Neural Information Processing Systems* 30 (2017) (vid. pág. 27).

Vollmer, Michael y Klaus-Peter Möllmann. *Infrared thermal imaging: Fundamentals, research and applications*. John Wiley & Sons, 2018 (vid. págs. 20, 22-24).

Williams Jr, George M. «Optimization of eyesafe avalanche photodiode lidar for automobile safety and autonomous navigation systems». En: *Optical Engineering* 56.3 (2017), págs. 031224-031224 (vid. pág. 13).

Xiang, Yuchen et al. «Hyperspectral Image Restoration and Super-resolution with Physics-Aware Deep Learning for Biomedical Applications». En: *arXiv preprint arXiv:2503.02908* (2025) (vid. pág. 29).

Xie, Enze et al. «SegFormer: Simple and efficient design for semantic segmentation with transformers». En: *Advances in Neural Information Processing Systems* 34 (2021), págs. 12077-12090 (vid. pág. 28).

Yang, Jinghe, Mingming Gong y Ye Pu. «Physics-Informed Knowledge Transfer for Underwater Monocular Depth Estimation». En: *European Conference on Computer Vision*. Springer. 2024, págs. 449-465 (vid. pág. 29).

Yang, Lihe et al. «Depth anything v2». En: *Advances in Neural Information Processing Systems* 37 (2024), págs. 21875-21911 (vid. págs. 35, 49).

Yang, Lihe et al. «Depth anything: Unleashing the power of large-scale unlabeled data». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 10371-10381 (vid. págs. 19, 28).

Yellin, Florence et al. «Concurrent Band Selection and Traversability Estimation from Long-Wave Hyperspectral Imagery in Off-Road Settings». En: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, págs. 7483-7492 (vid. pág. 60).

Zhang, Chuanqi et al. «Monocular Absolute Depth Estimation from Motion for Small Unmanned Aerial Vehicles by Geometry-Based Scale Recovery». En: *Sensors (Basel, Switzerland)* 24.14 (2024), pág. 4541 (vid. pág. 18).

Zuo, Chao et al. «Phase shifting algorithms for fringe projection profilometry: A review». En: *Optics and Lasers in Engineering* 109 (2018), págs. 23-59 (vid. pág. 18).