SEGMENTACIÓN DE MATERIALES A PARTIR DE IMÁGENES RGB USANDO ARQUITECTURAS DE TRANSFORMADORES DE VISIÓN E INTEGRACIÓN DE INFORMACIÓN MULTIESPECTRAL

NELSON FABIAN PEREZ PEREZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2024

SEGMENTACIÓN DE MATERIALES A PARTIR DE IMÁGENES RGB USANDO ARQUITECTURAS DE TRANSFORMADORES DE VISIÓN E INTEGRACIÓN DE INFORMACIÓN MULTIESPECTRAL

NELSON FABIAN PEREZ PEREZ

Trabajo de Grado para optar al título de
Ingeniero de Sistemas

Director:
Hoover Fabián Rueda Chacón
PhD. en Ingeniería Eléctrica y Computación

Codirector:
Brayan Esneider Monroy Chaparro
MSc. en Ingeniería de Sistemas e Informática

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2024

DEDICATORIA

A mi padre, Miguel, por todas sus enseñanzas y amor a lo largo de mi vida
A mi madre, Anyul, por su berraquera y amor incondicional. Yo juego para vos mamá.
A mi hermana, Angie, por su amor y por ser una fuente de inspiración y admiración.
A mi hermana, Yuli, por su resiliencia ejemplar y por el amor que me ha brindado.
A mis sobrinos, María y Martín, porque ustedes son el futuro de la familia.
A Pocholo, por acompañarme silenciosamente en mi escritorio todas las noches.

Este trabajo no hubiera sido posible sin su amor, compañía y apoyo incondicional. Se lo debo todo a ustedes, y cada logro en mi vida es completamente suyo.

AGRADECIMIENTOS

Agradezco a mis padres, los seres más valientes que he conocido, cuya fortaleza y amor han sido la base de todo lo que soy.

A mis hermanas, por estar siempre a mi lado con empatía, brindándome apoyo y guía en cada paso de este recorrido.

A toda mi familia, cuyo amor constante ha sido mi cimiento y me ha formado para ser quien soy hoy.

A mi director, Hoover, por ser un faro de orientación académica, por su confianza en mí y por permitirme explorar las ideas más atrevidas con libertad y creatividad.

A Brayan y Jorge, por su apoyo en la construcción de esta tesis. La admiración ha sido un motor clave en este proceso.

A Jhon y Karen, mis primeros mentores, por la paciencia con la que me introdujeron al mundo de la investigación.

A mis amigos, porque gracias a ellos este camino ha sido mucho más llevadero, por creer en mí en momentos de locura, y por compartir tantos buenos momentos que han hecho de esta camino algo inolvidable.

CONTENIDO

LISTA DE FIGURAS

pág.

LISTA DE CUADROS

pág.

RESUMEN

TÍTULO: SEGMENTACIÓN DE MATERIALES A PARTIR DE IMÁGENES RGB USANDO ARQUITECTURAS DE TRANSFORMADORES DE VISIÓN E INTEGRACIÓN DE INFORMACIÓN MULTIESPECTRAL[*]

AUTOR: NELSON FABIAN PEREZ PEREZ[**]

PALABRAS CLAVE: *Transformers* de visión, Aprendizaje multimodal, Imágenes espectrales, Modalidad faltante.

DESCRIPCIÓN:

La segmentación de materiales en imágenes RGB es una tarea desafiante debido a la complejidad de las texturas y la variabilidad de las condiciones de iluminación de los materiales. Aunque la información espectral puede mejorar significativamente esta tarea, su uso está limitado por la escasez de sensores espectrales en aplicaciones del mundo real. En este trabajo, presentamos un novedoso enfoque que integra eficientemente información espectral en un modelo de segmentación basado en *transformers*, manteniendo la capacidad de operar solo con imágenes RGB durante la inferencia. Nuestro modulo propuesto, denominado *Adaptive Spectral Prompts* (ASP), incorpora *prompts* espectrales adaptativos que se ajustan dinámicamente durante el entrenamiento, permitiendo al modelo aprovechar la riqueza de la información espectral sin depender de ella en la inferencia. Ademas, implementamos una estrategia de *modality dropout* para mejorar la robustez del modelo ante la ausencia de datos espectrales. Evaluamos exhaustivamente nuestro método en el dataset LIB-HSI, logrando un rendimiento significativo, con una precisión del $88{,}36\,\%$ y un IoU promedio por clase de $53{,}28\,\%$, superando significativamente a los métodos existentes. Nuestros experimentos demuestran la eficacia de ASP para integrar información multimodal de manera eficiente, mejorando la segmentación de materiales incluso en escenarios con modalidad faltante.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Hoover Fabián Rueda Chacón. Codirector: Brayan Esneider Monroy Chaparro.

ABSTRACT

TITLE: MATERIAL SEGMENTATION FROM RGB IMAGES USING VISION TRANSFORMER ARCHITECTURES AND MULTISPECTRAL INFORMATION*

AUTHOR: NELSON FABIAN PEREZ PEREZ**

KEYWORDS: Vision Transformers, Multimodal Learning, Spectral Images, Missing Modality.

DESCRIPTION:

Material segmentation in RGB images is a challenging task due to the complexity of textures and the variability of lighting conditions for materiales. Although spectral information can significantly improve this task, its use is limited by the scarcity of spectral sensors in real-world applications. In this work, we present a novel approach that efficiently integrates spectral information into a transformer-based segmentation model, while maintaining the ability to operate with RGB images during inference. Our architecture, called Adaptive Spectral Prompts (ASP), incorporates adaptive spectral prompts that dynamically adjust during training, allowing the model to leverage the richness of spectral information without depending on it during inference. We implement a modality dropout strategy to improve the model robustness in the absence of spectral data. We exhaustively evaluate our method on the LIB-HSI dataset, achieving state-of-the-art performance with an accuracy of $88{,}36\,\%$ and an average IoU per class of $53{,}28\,\%$, significantly outperforming existing methods. Our experiments demonstrate the effectiveness of ASP in efficiently integrating multimodal information, improving material segmentation even in scenarios with missing modalities.

---

* Bachelor Thesis

** Faculty of Physical-Mechanical Engineering. School of Computer Science. Advisor: Hoover Fabián Rueda Chacón. Co-advisor: Brayan Esneider Monroy Chaparro

1. INTRODUCCIÓN

La inteligencia artificial (IA) ha transformado radicalmente el desarrollo de algoritmos para tareas especializadas,[1,2] observándose un aumento significativo en su rendimiento en comparación con métodos tradicionales.[3] En particular, una de las áreas que más ha tenido impacto por la IA es el área de la visión por computadora, que se encarga de extraer información útil a partir de imágenes digitales.[4] A pesar de que las técnicas computacionales tradicionales establecieron una base sólida, su capacidad para manejar la complejidad y el volumen de estos datos era limitada, principalmente por la necesidad de seleccionar manualmente características relevantes y aplicar diversas heurísticas, lo cual era un proceso laborioso y propenso a errores. La llegada de métodos avanzados de aprendizaje profundo ha permitido superar muchas de estas limitaciones, logrando extraer información más detallada y precisa de grandes volúmenes de datos.[5] Este campo abarca diversas aplicaciones,

---

1 Yash Raj Shrestha, Shiko M Ben-Menahem y Georg Von Krogh. "Organizational decision-making structures in the age of artificial intelligence". En: *California Management Review* 61.4 (2019), págs. 66-83.

2 Iain M Cockburn, Rebecca Henderson y Scott Stern. *The impact of artificial intelligence on innovation*. Vol. 24449. National Bureau of Economic Research Cambridge, MA, USA, 2018.

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

4 Mahmoud Hassaballah y Ali Ismail Awad. *Deep learning in computer vision: principles and applications*. CRC Press, 2020.

5 Niall O'Mahony et al. "Deep learning vs. traditional computer vision". En: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. 2020, págs. 128-144.

desde la clasificación[6] hasta la detección[7] y restauración de imágenes,[8] siendo la segmentación de imágenes una de las más destacadas.[9,10] La segmentación implica clasificar los píxeles de una imagen en regiones homogéneas para facilitar su análisis, reconocimiento, y toma de decisiones.

Dentro de la segmentación de imágenes, la segmentación de materiales juega un rol crucial ya que permite identificar los componentes de una escena según los materiales presentes, como metal, vidrio, concreto, entre otros.[11] A pesar de los avances en algoritmos de IA para esta tarea usando imágenes RGB (imágenes que capturan información en tres canales de color: rojo, verde y azul, dentro del espectro visible), persisten limitaciones significativas en su desempeño debido a la confianza exclusiva en información RGB, lo que puede llevar a clasificaciones incorrectas por fenómenos como el metamerismo, la oclusión, o diferencias en textura.[12] Además, la falta de exploración en arquitecturas de IA más avanzadas limita el entendimiento y la precisión en la clasificación de materiales. Es por esto que se ve una clara diferencia

---

6 Jia Deng et al. "Imagenet: A large-scale hierarchical image database". En: *2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Ieee. 2009, págs. 248-255.

7 Zhong-Qiu Zhao et al. "Object detection with deep learning: A review". En: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (2019), págs. 3212-3232.

8 Jingyun Liang et al. "Swinir: Image restoration using swin transformer". En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 1833-1844.

9 German Ros et al. "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2016, págs. 3234-3243.

10 Yanming Guo et al. "A review of semantic segmentation using deep neural networks". En: *International Journal of Multimedia Information Retrieval* 7 (2018), págs. 87-93.

11 Edward H Adelson. "On seeing stuff: the perception of materials by humans and machines". En: *Human Vision and Electronic Imaging VI*. Vol. 4299. SPIE. 2001, págs. 1-12.

12 Yupeng Liang et al. "Multimodal material segmentation". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 19800-19808.

en el rendimiento al evaluar esta tarea contra sus contrapartes como segmentación semántica o detección de objetos las cuales encuentran un rendimiento satisfactorio en el estado del arte.[13,14]

Frente a este escenario, investigaciones recientes han intentado superar estas barreras, enfocándose en la ampliación de conjuntos de datos,[15] el desarrollo de arquitecturas más complejas[16] y la integración de diferentes modalidades de datos.[17] Una modalidad que destaca es la modalidad espectral, que utiliza información espectral para la segmentación, aprovechando el hecho de que las diferencias en los componentes que constituyen cada material presentan una interacción particular en el espectro electromagnético, también llamándose una firma espectral representativa, lo que permite la discriminación basada en las propiedades del material. Las imágenes espectrales, que incluyen información a lo largo de múltiples longitudes de onda, permiten una comprensión más profunda y precisa de los materiales presentes en una escena.[18,19] Sin embargo, adquirir esta información no es una tarea fácil

---

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

[14] 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.

[15] Paul Upchurch y Ransen Niu. "A dense material segmentation dataset for indoor and outdoor scene parsing". En: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2022, págs. 450-466.

[16] Tete Xiao et al. "Unified perceptual parsing for scene understanding". En: *Proceedings of the European Conference On Computer Vision (ECCV)*. Springer. 2018, págs. 418-434.

[17] Liang et al., "Multimodal material segmentation", ver n. 12.

[18] S Adarsh et al. "Performance comparison of Infrared and Ultrasonic sensors for obstacles of different materials in vehicle/robot navigation applications". En: *IOP Conference Series: Materials Science and Engineering*. Vol. 149. 1. IOP publishing. 2016, pág. 012141.

[19] Jeff W Lichtman y José-Angel Conchello. "Fluorescence microscopy". En: *Nature Methods* 2.12 (2005), págs. 910-919.

ni económica, y su uso todavía está restringido a laboratorios. En contraste, las imágenes RGB son de uso general, y los sensores de color son de muy fácil acceso; no obstante, extraer información de materiales utilizando solo tres canales es un desafío. A pesar de estos esfuerzos, las soluciones actuales aún no abordan de manera efectiva los desafíos inherentes a la segmentación de materiales, ya sea por su complejidad, la ineficiencia de las arquitecturas, o por limitarse a modalidades de datos costosas y poco prácticas.

A nuestro conocimiento, no se ha reportado en el estado del arte un enfoque que incorpore información espectral en la arquitectura de la red neuronal usando la reflectancia de múltiples longitudes de onda que proporcionan información más detallada de la escena para una clasificación guiada a partir de las firmas espectrales representativas de cada material. Además, que utilice las arquitecturas modernas de inteligencia artificial como los transformadores de visión (ViT, del inglés *Vision Transformers*),[20] que han revolucionado el rendimiento en otras tareas de visión por computadora. Las incrustaciones *embeddings* son un componente clave en estas arquitecturas, permitiendo la integración efectiva de información a diferentes niveles de abstracción y mejorando la capacidad del modelo para capturar patrones complejos y relaciones en los datos.[21,22] Este enfoque es particularmente útil en la segmentación de materiales donde la información detallada de múltiples longitudes de onda puede ser incrustada para mejorar la precisión y la fiabilidad del modelo, todo esto, manteniendo la facilidad de uso en el día a día con cámaras RGB, que

---

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

[21] Bowen Cheng, Alex Schwing y Alexander Kirillov. “Per-pixel classification is not all you need for semantic segmentation”. En: *Advances in Neural Information Processing Systems* 34 (2021), págs. 17864-17875.

[22] Andrew Jaegle et al. “Perceiver IO: A General Architecture for Structured Inputs & Outputs”. En: *International Conference on Learning Representations*. 2022.

son económicas y están ampliamente disponibles en comparación con las cámaras espectrales, que son más costosas y menos accesibles.

En el presente trabajo se desarrolló un novedoso algoritmo de segmentación de materiales en imágenes RGB, basado en aprendizaje profundo con *transformers* de visión, utilizando como incrustación la información espectral de cada tipo de material para la clasificación de los píxeles. En esta arquitectura, la imagen RGB fue la entrada del *transformer* de visión, mientras que la información espectral de los materiales de la escena se incrustó en la estructura del *transformer* en la etapa de entrenamiento. La metodología propuesta incluyó el establecimiento de las bases de datos de materiales y espectrales, la construcción de la arquitectura del *transformer* de visión para la segmentación de imágenes RGB según su material, y la implementación de una estrategia robusta para manejar escenarios de modalidad faltante, permitiendo al modelo operar eficazmente incluso cuando la información espectral no está disponible durante la inferencia. Se realizaron extensos estudios de ablación para validar cada componente del método propuesto. Además, se llevaron a cabo comparaciones exhaustivas con métodos del estado del arte, logrando superar el rendimiento de estos y estableciendo un nuevo estándar en la tarea de segmentación de materiales. Finalmente, se validó la capacidad de generalización del modelo utilizando escenas reales capturadas por los autores, demostrando la robustez y aplicabilidad del método propuesto en entornos no controlados.

2. OBJETIVOS

Objetivo general

Desarrollar y validar un algoritmo de segmentación de materiales basado en arquitecturas de *transformers* de visión que integre información espectral sobre imágenes de color (RGB).

Objetivos específicos

1. Identificar y seleccionar bases de datos de imágenes espectrales y de color (RGB) adecuadas para el entrenamiento y prueba del algoritmo, asegurando una amplia representación de materiales y condiciones de iluminación.

2. Diseñar una arquitectura de *transformer* de visión que integre información espectral y de color (RGB) de una escena para segmentarla en términos de los materiales que la componen.

3. Implementar en Python la arquitectura de *transformer* de visión diseñada para la segmentación de los materiales de una escena.

4. Evaluar el desempeño del algoritmo desarrollado mediante métricas de rendimiento estándar en el área de segmentación.

5. Validar cualitativamente el algoritmo desarrollado sobre un conjunto de imágenes de color (RGB) adquiridas con una cámara disponible en dispositivos electrónicos de consumo.

3. MARCO DE REFERENCIA

3.1. Segmentación de materiales

La segmentación de materiales es un proceso que busca identificar los distintos materiales presentes en una escena, mediante la división de la imagen en múltiples regiones homogéneas, cada una correspondiendo a la etiqueta de un material específico, como metal, vidrio o madera.[23] Investigaciones recientes[24] han resaltado la importancia de diferenciar entre cosas (objetos con una definición clara) y "elementos" o *stuff* (materiales o texturas sin una forma definida), subrayando que la capacidad para reconocer y diferenciar materiales es tan crucial como la identificación de objetos en una escena. Esta distinción no es solo esencial para una comprensión detallada de las escenas, sino que también tiene implicaciones prácticas significativas, como mejorar la autonomía de robots,[25] la simulación acústica,[26] y las aplicaciones de realidad mixta consciente del contexto,[27] las cuales dependen de un reconocimiento preciso del entorno.

La segmentación de materiales enfrenta desafíos únicos, dado que los materiales pueden variar ampliamente en apariencia y textura, lo que puede confundir los algoritmos de detección. Fenómenos como el metamerismo, ilustrado en la Figura 1, donde

---

[23] Upchurch y Niu, ver n. 15.

[24] Adelson, ver n. 11.

[25] Adarsh et al., ver n. 18.

[26] Anurag Arnab et al. "Joint Object-Material Category Segmentation from Audio-Visual Cues". En: *Proceedings of the British Machine Vision Conference (BMVC)*. 2015.

[27] Long Chen et al. "Context-Aware Mixed Reality: A Learning-Based Framework for Semantic-Level Interaction". En: *Computer Graphics Forum*. Vol. 39. 1. Wiley Online Library. 2020, págs. 484-496.

materiales con diferentes propiedades espectrales parecen idénticos bajo ciertas condiciones de iluminación, complican aún más la tarea debido a sus apariencias similares, llevando a clasificaciones incorrectas.[28] Además, el desbalance de clases es un problema frecuente, ya que algunos materiales son mucho más comunes que otros en el día a día, y por tanto estarán mucho más representados en los conjuntos de datos, llevando a clasificar con mas dificultad las clases menos representadas.

Figura 1. Las superficies con texturas similares pueden estar compuestas de materiales diferentes. Estos objetos son de tela, plástico y papel, de izquierda a derecha. Tomado de.[29]

Las etiquetas de los materiales se escogen para imitar la percepción humana, seleccionando materiales distinguidos por propiedades útiles y físicas reconocibles. Algunos estudios han propuesto esquemas de etiquetado jerárquicos, donde ciertas etiquetas son subcategorías de otras, como diferentes tipos de madera o tejidos.[30] En la Figura 2 se muestra un ejemplo de una segmentación con sus respectivas etiquetas.

---

[28] David H. Foster et al. “Frequency of metamerism in natural scenes”. En: *J. Opt. Soc. Am. A* 23.10 (2006), págs. 2359-2372.

[29] Ce Liu et al. “Exploring features in a bayesian framework for material recognition”. En: *2010 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, págs. 239-246.

[30] Gabriel Schwartz y Ko Nishino. “Recognizing material properties from images”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.8 (2019), págs. 1981-1995.

Figura 2. Ejemplo de segmentación de materiales aplicada a una imagen RGB, ilustrando la identificación y clasificación de diferentes materiales como metal, plástico, concreto y caucho. Tomado de.[31]

Existen dos enfoques predominantes en la segmentación de materiales: los basados en imágenes RGB (*Red, Green and Blue*, por sus siglas en inglés) y aquellos que utilizan modalidades de datos no-RGB. Los métodos basados en RGB se han centrado en la construcción de extensos conjuntos de datos.[32,33] Aunque estos conjuntos de datos inicialmente presentaban escasez en el etiquetado, es decir, gran cantidad de pixeles sin una etiqueta asociada, esfuerzos recientes han culminado en la creación de colecciones de datos densamente etiquetados, que han permitido la segmentación automática a través de redes neuronales convolucionales con un notable éxito.[34] Por otro lado, los enfoques basados en datos no-RGB han explorado

---

[31] Paul Upchurch y Ransen Niu. "A dense material segmentation dataset for indoor and outdoor scene parsing". En: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2022, págs. 450-466.

[32] Schwartz y Nishino, ver n. 30.

[33] Sean Bell et al. "Material recognition in the wild with the materials in context database". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2015, págs. 3479-3487.

[34] Upchurch y Niu, ver n. 31.

el uso de la Función de Distribución de Reflectancia Bidireccional (BRDF)[35] y la polarización,[36] logrando mejoras en rendimiento, pero a su vez introduciendo barreras para su uso cotidiano debido a su complejidad de implementación y requerimientos de hardware específico costoso. La información espectral se destaca como una modalidad prometedora para la clasificación de materiales, ya que proporciona firmas representativas para cada tipo de material, y ha demostrado que con su uso se mejora el rendimiento respecto a la segmentación basada en solo RGB.[37]

---

[35] Jia Xue et al. “Differential angular imaging for material recognition”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017, págs. 764-773.

[36] Liang et al., “Multimodal material segmentation”, ver n. 12.

[37] Nariman Habili et al. “A hyperspectral and RGB dataset for building façade segmentation”. En: *European Conference on Computer Vision*. Springer. 2022, págs. 258-267.

3.2. Imágenes espectrales

Las imágenes espectrales representan una rica fuente de información más allá de lo que el ojo humano o las cámaras RGB convencionales pueden capturar, al registrar la intensidad de la luz en una amplia gama de longitudes de onda para cada punto de la imagen.[38] La característica fundamental de este enfoque radica en que cada material, debido a su composición química y estructura física específica, interactúa de manera única con la luz, absorbiendo y reflejando diferentes longitudes de onda.[39] Esta interacción distintiva queda plasmada en lo que se conoce como firma espectral. La capacidad de las imágenes espectrales para capturar estos patrones de manera detallada permite la identificación precisa de los materiales presentes en una escena, como se ilustra en la Figura 3. Estas firmas son esenciales para distinguir materiales con gran precisión en numerosas aplicaciones, como el sensado remoto para la detección de minerales,[40] análisis de alimentos[41] y análisis de suelos en la agricultura de precisión.[42]

Específicamente, las imágenes espectrales se adquieren mediante sensores especializados capaces de discriminar entre una extensa gama de longitudes de onda. Existen principalmente dos tipos: las imágenes multiespectrales, que capturan infor-

---

[38] Yuval Garini, Ian T Young y George McNamara. "Spectral imaging: principles and applications". En: *Cytometry part a: The Journal of the International Society for Analytical Cytology* 69.8 (2006), págs. 735-747.

[39] Adarsh et al., ver n. 18; Lichtman y Conchello, ver n. 19.

[40] Hojat Shirmard et al. "A review of machine learning in processing remote sensing data for mineral exploration". En: *Remote Sensing of Environment* 268 (2022), pág. 112750.

[41] Yuwei Liu, Hongbin Pu y Da-Wen Sun. "Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications". En: *Trends in Food Science & Technology* 69 (2017), págs. 25-35.

[42] Bing Lu et al. "Recent advances of hyperspectral imaging technology and applications in agriculture". En: *Remote Sensing* 12.16 (2020), pág. 2659.

mación en un número limitado de bandas, típicamente entre 3 y 10, y las imágenes hiperespectrales, que registran datos en cientos o incluso miles de bandas,[43] ofreciendo una resolución espectral detallada. Este nivel de detalle en la captura permite una percepción y análisis más profundos del entorno capturado.

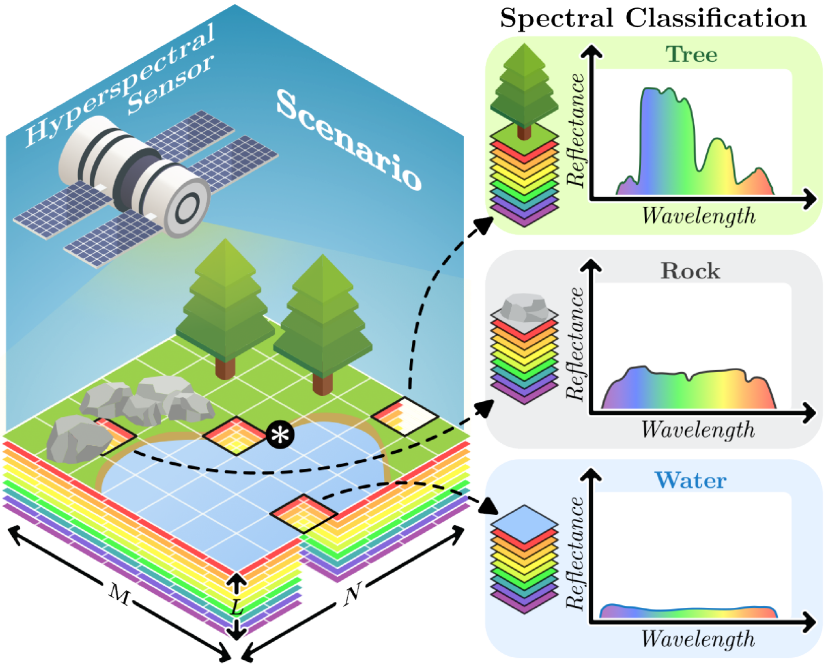

Figura 3. Descripción general de las imágenes espectrales. Una cámara espectral registra la reflectancia en forma de imágenes. Estas imágenes permiten la identificación de diferentes materiales, como la vegetación *(tree)*, las rocas *(rock)* o el agua *(water)*, a través de sus firmas espectrales representativas. Tomada de [43].

---

[43] Jorge Bacca, Emmanuel Martinez y Henry Arguello. "Computational spectral imaging: A contemporary overview". En: *JOSA A* 40.4 (2023), págs. C115-C125.

3.3. *Transformers* de visión

Los *transformers*[44] han revolucionado el campo del procesamiento del lenguaje natural (NLP) gracias a su estructura innovadora que favorece la paralelización, la escalabilidad y la capacidad de generalización, sin introducir sesgos debido a su arquitectura. Esta eficiencia ha propiciado su adaptación en el ámbito de la visión por computadora, donde han emergido como arquitecturas libres de convoluciones[45,46] que han demostrado un rendimiento superior en diversas tareas como clasificación,[47] segmentación[48] y detección.[49] En la Figura 4 se ilustra la arquitectura de el *encoder* de un *transformer*.

El corazón de los *transformers* es el mecanismo de atención *(attention)*, que permite capturar dependencias y relaciones a largo plazo dentro de los datos. Este mecanismo asigna un conjunto de puntuaciones de atención a cada elemento de entrada, basándose en la comparación de pares de elementos mediante las operaciones:

---

[44] Ashish Vaswani et al. “Attention is all you need”. En: *Advances in Neural Information Processing Systems* 30 (2017).

[45] Namyup Kim et al. “Restr: Convolution-free referring image segmentation using transformers”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 18145-18154.

[46] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 10012-10022.

[47] Dosovitskiy et al., ver n. 20.

[48] Jitesh Jain et al. “Oneformer: One transformer to rule universal image segmentation”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 2989-2998.

[49] Xizhou Zhu et al. “Deformable detr: Deformable transformers for end-to-end object detection”. En: *arXiv preprint arXiv:2010.04159* (2020).

[50] Ashish Vaswani et al. “Attention is all you need”. En: *Advances in Neural Information Processing Systems* 30 (2017).

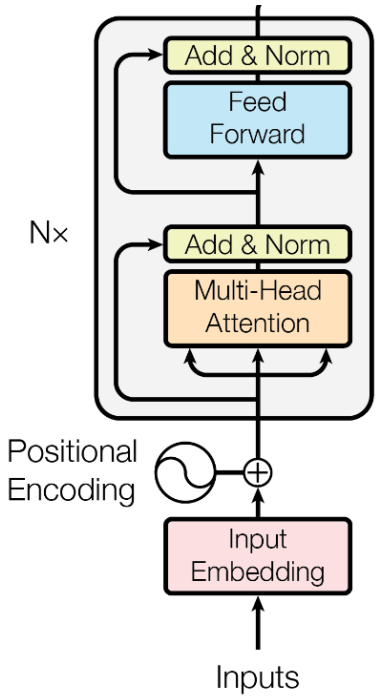

Figura 4. Arquitectura del *encoder* de un *transformer* mostrando el proceso secuencial desde el *embedding* de entrada y codificación posicional hasta la atención multi-cabeza y la red *feed-forward*. Adaptada de.[50]

consulta *(query)*, clave *(key)* y valor *(value)*. Estos términos, aunque pueden evocar conceptos de bases de datos, se refieren en realidad a transformaciones vectoriales: la consulta *(query)*, representa el vector de información que se está analizando, la clave *(key)* es un vector que se compara con la consulta *(query)* para determinar su relevancia, y el valor *(value)* es el vector de información que se ponderará según la atención calculada. La atención, en este contexto, se refiere a la capacidad del modelo para enfocarse en partes específicas y relevantes de los datos de entrada al procesar cada elemento, permitiendo al modelo centrarse en las partes más relevantes de los datos para cada tarea específica. Esta operación es potenciada por el concepto de atención multi-cabeza *(multi-head attention)*, donde cada cabeza *(head)* es un conjunto independiente de transformaciones lineales aplicadas a las consultas, claves y valores que permite al modelo procesar y atender en distintas partes de

la entrada en el mismo nivel de abstracción de la red neuronal, mejorando así la capacidad para enfocarse en múltiples aspectos de los datos de manera simultánea. Cada cabeza de atención descubre relaciones únicas entre los elementos de entrada, enriqueciendo la comprensión global del modelo sobre la estructura y el contenido.
Los *transformers* están compuestos por módulos que procesan los datos en etapas sucesivas, añadiendo capas de análisis y refinamiento. La combinación de estos módulos, una vez entrenados, confiere una extraordinaria flexibilidad y numerosos grados de libertad, permitiendo que los *transformers* se ajusten eficazmente a cualquier tipo de dato o tarea.[51] Esta escalabilidad se evidencia al entrenar con vastos conjuntos de datos, observándose una mejora en el rendimiento, proporcional al aumento del volumen de datos.[52,53] Siguiendo esta línea, investigaciones recientes[54] han validado su eficacia incluso con conjuntos de datos más reducidos mediante técnicas de ajuste fino *(fine-tuning)*, expandiendo su aplicabilidad.
En el contexto de la visión por computadora, se han propuesto múltiples arquitecturas de *transformers* de visión adaptados a una gran variedad de tareas. Ejemplos notables incluyen el *Vision Transformer* (ViT)[55] y el *Data-efficient Image Transformer* (DeiT),[56] diseñados originalmente para clasificación de imágenes. Estos enfoques se

---

[51] Hugo Touvron et al. “Going deeper with image transformers”. En: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, págs. 32-42.

[52] Jared Kaplan et al. “Scaling laws for neural language models”. En: *arXiv preprint arXiv:2001.08361* (2020).

[53] Xiaohua Zhai et al. “Scaling vision transformers”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 12104-12113.

[54] Zhixiang Wei et al. “Stronger Fewer & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 28619-28630.

[55] Dosovitskiy et al., ver n. 20.

[56] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. En:

han extendido a otras aplicaciones incluyendo, la clasificación de vídeo,[57] la segmentación semántica,[58] (ilustrada en la Figura 5), y la detección de objetos.[59] Un enfoque común en el procesamiento de imágenes con *transformers* es dividir las imágenes en pequeños fragmentos llamados "parches". Luego, estos parches se convierten en representaciones numéricas *(tokens)* a las que se les añade información sobre su posición en la imagen *(embeddings)*. Esto permite que el modelo mantenga la estructura espacial de la imagen, pudiendo tratarla como una secuencia de datos similar a cómo se procesan las palabras en el procesamiento de texto.

Recientemente, se ha intensificado el esfuerzo por adaptar los *transformers* de visión para su uso en hardware de consumo,[60] buscando hacer accesibles estas potentes herramientas de aprendizaje profundo para aplicaciones en tiempo real en dispositivos con capacidades computacionales más limitadas. Estas optimizaciones en el diseño y la implementación de algoritmos eficientes han permitido que los *transformers* ofrezcan un rendimiento excepcional no solo en servidores de alto rendimiento, sino también en ambientes computacionales más restringidos, abriendo nuevas posibilidades para su implementación en una amplia gama de aplicaciones.

International Conference on Machine Learning. PMLR. 2021, págs. 10347-10357.

[57] Anurag Arnab et al. “Vivit: A video vision transformer”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 6836-6846.

[58] Robin Strudel et al. “Segmenter: Transformer for semantic segmentation”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 7262-7272.

[59] Nicolas Carion et al. “End-to-end object detection with transformers”. En: *European Conference on Computer Vision*. Springer. 2020, págs. 213-229.

[60] Yanyu Li et al. “Efficientformer: Vision transformers at mobilenet speed”. En: *Advances in Neural Information Processing Systems* 35 (2022), págs. 12934-12949.

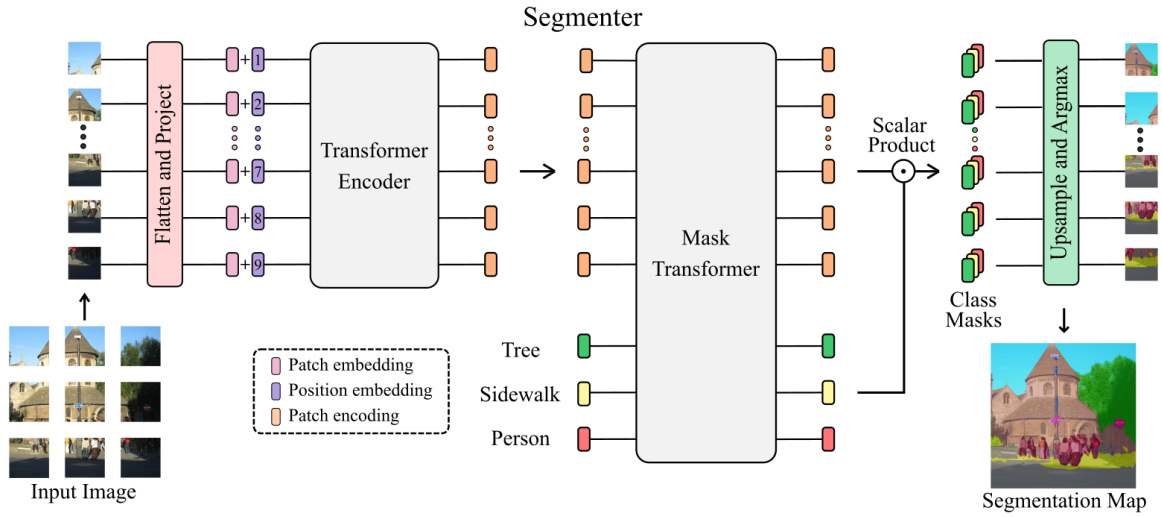

Figura 5. Ejemplo de arquitectura basada en *transformers* para la segmentación de imágenes, mostrando el flujo desde la entrada de la imagen hasta la generación del mapa de segmentación final. Tomada de.[61]

3.4. Aprendizaje multimodal

El aprendizaje multimodal se refiere a la capacidad de los sistemas de inteligencia artificial para procesar y combinar información proveniente de múltiples fuentes o modalidades, como texto, imágenes, audio u otro tipo de modalidades.[62] Este enfoque busca mimetizar la forma en que los humanos integramos naturalmente información de diferentes sentidos para comprender mejor nuestro entorno. En el campo de la visión por computadora, el aprendizaje multimodal ha demostrado ser particularmente efectivo en diversas tareas. Por ejemplo, la combinación de texto e imágenes ha llevado a avances en áreas como la generación de descripciones de imágenes.[63] Asimismo, la fusión de información RGB con datos de profundidad

---

[62] Jiquan Ngiam et al. “Multimodal deep learning”. En: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, págs. 689-696.

[63] Pan Lu et al. “Learn to explain: Multimodal reasoning via thought chains for science question answering”. En: *Advances in Neural Information Processing Systems* 35 (2022), págs. 2507-2521;

(RGB-D) ha mejorado notablemente el rendimiento en tareas como la detección de objetos y la segmentación semántica.[64,65] En el ámbito del procesamiento de video, la integración de información visual y auditiva ha permitido desarrollar sistemas más robustos para la detección de eventos y la comprensión de escenas.[66,67] Los modelos multimodales pueden estructurarse de diversas formas, dependiendo de cómo se fusionan las diferentes modalidades de datos, como se ilustra en la Figura 6.

Penghao Wu y Saining Xie. "V?: Guided Visual Search as a Core Mechanism in Multimodal LLMs". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 13084-13094.

[64] Wei Gao et al. "Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection". En: *IEEE Transactions on Circuits and Systems for Video Technology* 32.4 (2021), págs. 2091-2106.

[65] Xiaokang Chen et al. "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation". En: *European Conference on Computer Vision*. Springer. 2020, págs. 561-577.

[66] Hassan Akbari et al. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text". En: *Advances in Neural Information Processing Systems* 34 (2021), págs. 24206-24221.

[67] Xiongkuo Min et al. "A multimodal saliency model for videos with high audio-visual correspondence". En: *IEEE Transactions on Image Processing* 29 (2020), págs. 3805-3819.

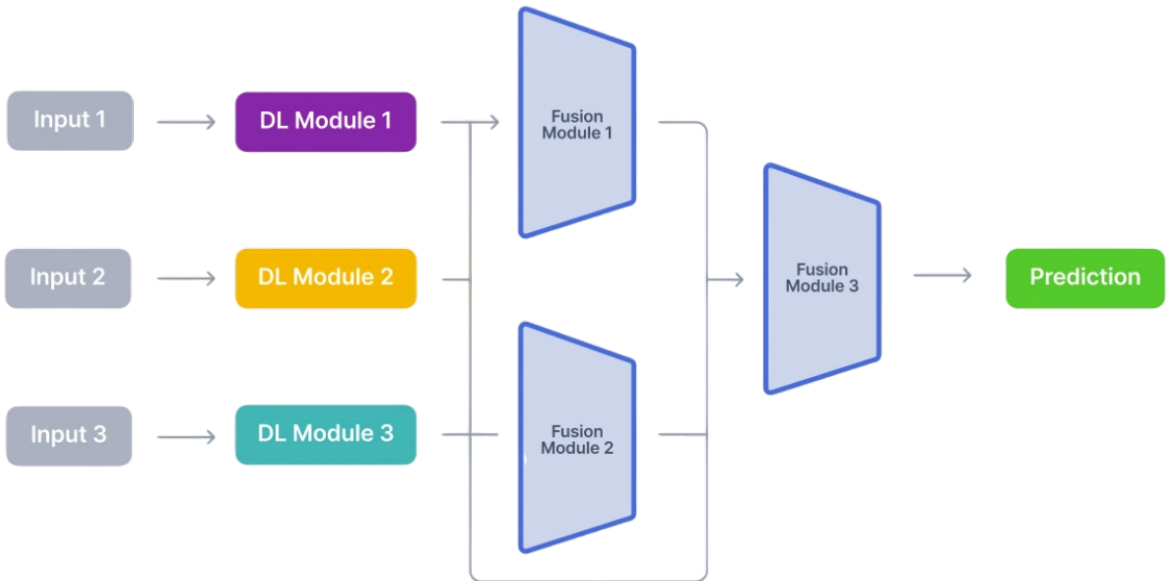

Figura 6. Ilustración conceptual de el aprendizaje multimodal, cada *(Input)* es una diferente modalidad que se procesa a través de módulos específicos. Tomada de[68]

Los *transformers* han demostrado ser particularmente efectivos en el aprendizaje multimodal, gracias a su capacidad para manejar secuencias de datos de longitud variable y capturar dependencias a largo plazo.[69] Una de las claves de su éxito radica en su capacidad para tokenizar diferentes tipos de datos, permitiendo que modalidades dispares como texto, imágenes o audio se representen en un espacio común de tokens.[70] Esta representación unificada facilita la interacción entre modalidades y permite que el modelo aprenda relaciones complejas entre ellas. En este contexto, avances recientes han expandido significativamente las capacidades multimodales de los *transformers*,[71] adaptando un modelo de lenguaje pre-entrenado unimodal a un modelo multimodal capaz de procesar tanto texto como imágenes. Otros en-

---

[69] Peng Xu, Xiatian Zhu y David A Clifton. "Multimodal learning with transformers: A survey". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023), págs. 12113-12132.

[70] Jaegle et al., ver n. 22.

[71] Abhimanyu Dubey et al. "The llama 3 herd of models". En: *arXiv preprint arXiv:2407.21783* (2024).

foques,[72,73] proponen arquitecturas *transformers* escalables diseñadas desde su concepción para el procesamiento multimodal, demostrando la capacidad de manejar hasta 12 modalidades diferentes simultáneamente.

Un aspecto crucial en el aprendizaje multimodal es la capacidad de manejar escenarios con modalidad faltante *(missing modality)*.[74] Este escenario se presenta cuando, durante la inferencia, no todas las modalidades utilizadas en el entrenamiento están disponibles. El manejo de la modalidad faltante es particularmente relevante en aplicaciones prácticas, ya que garantiza la robustez del modelo en situaciones del mundo real donde ciertas fuentes de datos pueden no estar accesibles.

Diversas técnicas se han propuesto para abordar este desafío. Una de estas técnicas implica el uso de modelos generativos para sintetizar las características faltantes a partir de las modalidades disponibles.[75] Alternativamente, algunos trabajos[76,77] han optado por entrenar modelos multimodales específicamente para escenarios con modalidades ausentes, mejorando así la robustez del sistema en condiciones de

---

[72] Siddharth Srivastava y Gaurav Sharma. "OmniVec2-A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 27412-27424.

[73] Yiyuan Zhang et al. "Meta-transformer: A unified framework for multimodal learning". En: *arXiv preprint arXiv:2307.10802* (2023).

[74] Renjie Wu, Hu Wang y Hsiang-Ting Chen. "A comprehensive survey on deep Multimodal Learning with Missing Modality". En: *arXiv [cs.CV]* (12 de sep. de 2024).

[75] Mengmeng Ma et al. "Smil: Multimodal learning with severely missing modality". En: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, págs. 2302-2310.

[76] Hu Wang et al. "Multi-modal learning with missing modality via shared-specific feature modelling". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 15878-15887.

[77] Yao Zhang et al. "mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation". En: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, págs. 107-117.

inferencia subóptimas. Un tercer paradigma, particularmente prometedor, se centra en el aprendizaje de espacios latentes compartidos que son inherentemente robustos a entradas faltantes,[78,79] esto permite a los modelos mantener un rendimiento consistente incluso cuando ciertas modalidades no están disponibles, utilizando decodificadores de tareas compartidas que pueden operar eficazmente con representaciones latentes parciales. Cada una de estas estrategias ofrece ventajas únicas y su aplicabilidad depende en gran medida del contexto específico de la aplicación y de las características de las modalidades involucradas.

3.5. *Prompt tuning*

El *prompt tuning* es una técnica emergente en el aprendizaje profundo, inicialmente desarrollada en el procesamiento del lenguaje natural (NLP), pero que recientemente ha cobrado importancia en la visión por computadora.[80] Esta técnica consiste en optimizar secuencias específicas de datos, conocidas como *prompts* o indicaciones, que se añaden a la entrada del modelo. Un *prompt* es esencialmente un conjunto de instrucciones o pistas que ayudan al modelo a entender mejor el contexto o la tarea que debe realizar.[81] En lugar de entrenar completamente un modelo para cada nueva tarea, el *prompt tuning* permite ajustar estos *prompts* de manera que el modelo preentrenado se oriente de forma más efectiva hacia la tarea deseada, mejorando su

---

[78] Mohammad Havaei et al. “Hemis: Hetero-modal image segmentation”. En: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, págs. 469-477.

[79] Yao-Hung Hubert Tsai et al. “Learning Factorized Multimodal Representations”. En: *International Conference on Learning Representations*. 2019.

[80] Menglin Jia et al. “Visual prompt tuning”. En: *European Conference on Computer Vision*. Springer. 2022, págs. 709-727.

[81] Pengfei Liu et al. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. En: *ACM Computing Surveys* 55.9 (2023), págs. 1-35.

desempeño sin necesidad de un reentrenamiento exhaustivo del modelo completo. En el contexto de los modelos de lenguaje, el *prompt tuning* ha demostrado ser una alternativa eficiente al ajuste fino tradicional, permitiendo adaptar grandes modelos pre-entrenados a tareas específicas con una fracción de los parámetros ajustables.[82] Este enfoque no solo reduce los requisitos computacionales en el entrenamiento, sino que también mejora la generalización del modelo a nuevas tareas.

La adaptación del *prompt tuning* a la visión por computadora presenta desafíos únicos debido a la naturaleza fundamentalmente diferente de los datos visuales en comparación con el texto. Sin embargo, investigaciones recientes han demostrado su potencial en tareas de visión.[83] En este contexto, los *prompts* pueden tomar la forma de *tokens* aprendibles que se concatenan con las características de la imagen o se integran en las capas del modelo. Asimismo, el *prompt tuning* ha demostrado ser efectivo en el aprendizaje multimodal, donde se combinan diferentes tipos de datos, como texto e imágenes. Khattak et al.[84] proponen un método que utiliza multiples *prompts* para guiar un modelo de visión y lenguaje en diversas tareas.

---

[82] Brian Lester, Rami Al-Rfou y Noah Constant. “The power of scale for parameter-efficient prompt tuning”. En: *arXiv preprint arXiv:2104.08691* (2021).

[83] Kaiyang Zhou et al. En: *International Journal of Computer Vision* 130.9 (2022), págs. 2337-2348.

[84] Muhammad Uzair Khattak et al. “Maple: Multi-modal prompt learning”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 19113-19122.

4. MÉTODO PROPUESTO

En este trabajo se desarrolló un método para la segmentación multimodal de materiales utilizando imágenes RGB e información espectral que se caracteriza por su robustez ante la ausencia de una modalidad, permitiendo la segmentación incluso cuando solo se dispone de información RGB. Este enfoque es capaz de aprovechar la información espectral cuando está disponible, lo que resulta en una mejora significativa del rendimiento en la tarea de segmentación de materiales. El enfoque propuesto se basa en una arquitectura tipo *transformer encoder-decoder* para predicción densa[85] , diseñada para procesar y fusionar eficientemente información de las modalidades RGB y espectral. El método implementado es robusto ante la ausencia de una modalidad, permitiendo la segmentación incluso cuando solo se dispone de información RGB.

La arquitectura desarrollada consta de una base *transformer* pre-entrenada para segmentación semántica RGB, que se complementa con un módulo de incrustación espectral y un módulo de adaptación denominado *Adaptive Spectral Prompts* (ASP). El módulo de incrustación espectral se diseñó específicamente para procesar eficientemente la información espectral, teniendo en cuenta sus características únicas y su relación con los datos RGB. Por su parte, ASP ajusta dinámicamente las representaciones espectrales aprendidas para optimizar la segmentación de materiales en diversos escenarios. Esta arquitectura híbrida aprovecha el poder de los modelos pre-entrenados mientras se adapta eficientemente a la tarea específica de segmentación multimodal de materiales.

---

[85] 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.

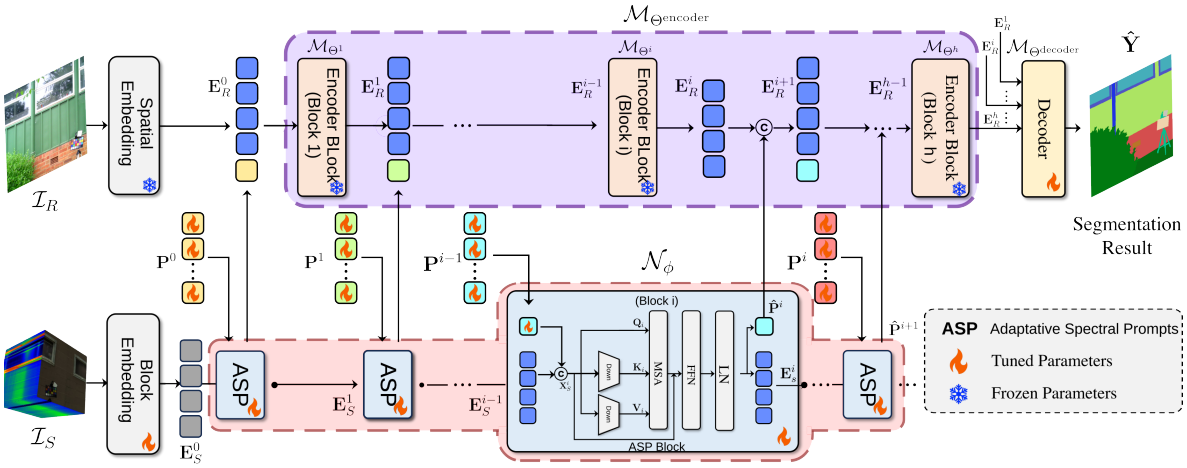

Figura 7. Arquitectura del método propuesto para la segmentación de materiales utilizando información RGB y espectral. El modelo incorpora un *encoder* con bloques de *Adaptive Spectral Prompts* (ASP) que procesan la información espectral, mientras que la rama superior procesa la información RGB. Los parámetros congelados (marcados con ❄) provienen de un modelo pre-entrenado, mientras que los parámetros ajustables (marcados con 🔥) se optimizan durante el entrenamiento. La arquitectura culmina en un *decoder* que genera el resultado final de segmentación.

Para nuestro método, las entradas al modelo de tipo *transformer encoder-decoder* $\mathcal{M}_\theta$ son la imagen espectral $\mathcal{I}_S \in \mathbb{R}^{H\times W\times B}$, donde H y W representan la altura y la anchura de la imagen, respectivamente, y B es el número de bandas espectrales, junto con $\mathcal{I}_R \in \mathbb{R}^{H\times W\times 3}$, correspondiente a la imagen RGB. Para integrar eficientemente la información espectral, se introduce un conjunto de *prompts*[86] aprendibles $\mathcal{P} = \{\mathbf{P}^0, \mathbf{P}^1, \ldots, \mathbf{P}^h\}$, donde h es el número de bloques del *encoder* $\mathcal{M}_{\theta\text{encoder}}$ y cada $\mathbf{P}^i \in \mathbb{R}^{l_i\times d_i}$, siendo l_i el número de *prompts* aprendibles y d_i la dimensión del bloque i del modelo $\mathcal{M}_\theta$. Cada uno de estos *prompts* $\mathbf{P}^i$ se inicializa de manera aleatoria usando la *kaiming* inicialización[87]

---

[86] Jia et al., ver n. 80.

[87] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". En: *Proceedings of the IEEE international conference on computer vision*. 2015, págs. 1026-1034.

El entrenamiento e inferencia del modelo propuesto $f(\cdot)$ en escenarios con ambas modalidades (RGB y espectral) sigue un flujo donde la imagen RGB $(\mathcal{I}_R)$ se procesa a través de el modelo pre-entrenado $\mathcal{M}_\theta$, mientras que la información espectral $\mathcal{I}_S$ se combina con los *prompts* $\mathcal{P}$ a través de los bloques ASP que interactúan con el modelo $\mathcal{M}_\theta$ así:

$$\hat{\mathbf{Y}} = f\left(\mathcal{M}_\theta(\mathcal{I}_R), \mathcal{N}_\phi(\mathcal{I}_S, \mathcal{P})\right), \tag{1}$$

donde $\mathcal{M}_\theta$ es el modelo pre-entrenado que procesa la imagen RGB, y $\mathcal{N}_\phi$ corresponde a los bloques *ASP*. El modelo propuesto $f(\cdot)$ genera una predicción $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W\times K}$ donde K es el número de clases de materiales que representa la segmentación final en la imagen.

En el escenario donde la modalidad espectral no está disponible, el modelo se mantiene funcional utilizando solo la información RGB y los *prompts* aprendibles. En este caso, se simplifica a:

$$\hat{\mathbf{Y}} = f\left(\mathcal{M}_\theta(\mathcal{I}_R), \mathcal{P}\right), \tag{2}$$

donde el conjunto de *prompts* $\mathcal{P}$ pasa directamente al modelo $f(\cdot)$ que encapsula $\mathcal{M}_\theta$ y que se encargará de concatenar los *prompts* a las características en los bloques intermedios de $\mathcal{M}_\theta$.

4.1. *Embeddings* espectrales por bloques

Tanto $\mathcal{M}_\theta(\cdot)$, como $\mathcal{N}_\phi(\cdot)$ son modelos tipo *transformer*[88] que procesan la informacion mediante secuencias de *tokens*. Para procesar eficientemente la información espectral de $\mathcal{I}_S$, se implementó un enfoque de construcción de *embeddings* por bloques que preserva la estructura espacial y espectral de los datos de entrada inspirados

[88] Vaswani et al., ver n. 50.

en.[89]

El objetivo es obtener una representación de tokens $\mathbf{E}_S \in \mathbb{R}^{n\times d}$, donde n es el número de *tokens* y d es la dimensión del espacio del *embedding*. Para esto, se redimensiona $\mathcal{I}_S$ en n bloques secuenciales 3D vectorizados $\hat{\mathcal{I}}_S \in \mathbb{R}^{n\times(p_h\cdot p_w\cdot p_c)}$, donde p_h, p_w y p_b son los tamaños del bloque en las dimensiones espaciales y espectrales, respectivamente. El número total de bloques es $n = (\frac{h}{p_h}) \cdot (\frac{w}{p_w}) \cdot (\frac{b}{p_b})$. Cada bloque $\mathbf{b}_k \in \mathbb{R}^{(p_h\cdot p_w\cdot p_c)}$, para $k = 1, 2, ..., n$, se proyecta a un vector de *embedding* $\mathbf{e}_i \in \mathbb{R}^d$ mediante una transformación lineal $\mathbf{W}_S \in \mathbb{R}^{(p_h\cdot p_w\cdot p_c)\times d}$ y para incorporar información posicional[90] se añade un *embedding* posicional aprendible $\boldsymbol{\alpha}_i \in \mathbb{R}^d$ a cada token:

$$\mathbf{e}_k = \mathbf{W}_S^\top \cdot \mathbf{b}_k + \boldsymbol{\alpha}_k, \qquad \mathbf{e}_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, n. \tag{3}$$

De esta forma, la colección de *embeddings* espectrales será $\mathbf{E}_S = \{\mathbf{e}_k^\top \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leq k \leq n\}$. En cuanto al procesamiento de la imagen RGB, se sigue el enfoque tradicional.[91] Sea $\mathcal{I}_R \in \mathbb{R}^{H\times W\times 3}$ la imagen RGB de entrada. Esta imagen $\mathcal{I}_R$ se redimensiona a n parches vectorizados $\hat{\mathcal{I}}_R \in \mathbb{R}^{n\times(p_{h'}\cdot p_{w'}\cdot 3)}$ donde cada parche $\mathbf{p}_k \in \mathbb{R}^{(p_{h'}\cdot p_{w'}\cdot 3)}$, específicamente $p_{h'}$ y $p_{w'}$ son las dimensiones del tamaño de cada parche. Cada parche se proyecta en un espacio latente d-dimensional mediante una transformación:

$$\hat{\mathbf{e}}_k = \mathcal{E}_\omega(\mathbf{p}_k), \qquad \mathbf{e}_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, n, \tag{4}$$

donde $\mathcal{E}_\omega(\cdot)$ representa la función de *embedding* del modelo $\mathcal{M}_\theta$ con parámetros

---

89 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.

90 Vaswani et al., ver n. 50.

91 Dosovitskiy et al., ver n. 20.

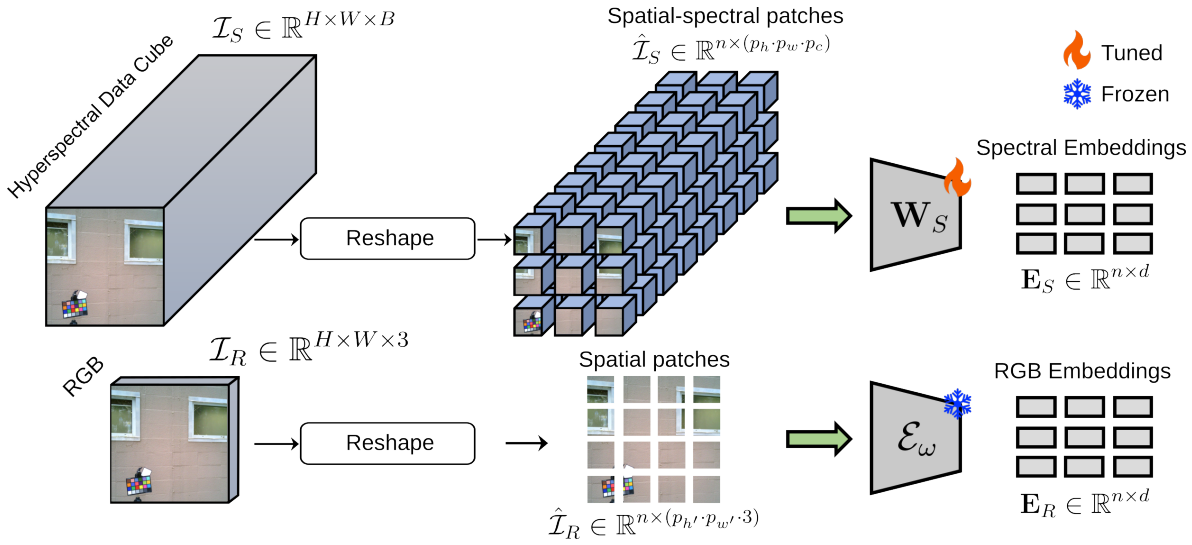

Figura 8. Esquema del proceso de generación de *embeddings* para las modalidades espectral y RGB. En la parte superior se muestra el proceso para los datos espectrales, donde se aplica un redimensionamiento seguido de una proyección lineal entrenable ($\mathbf{W}_S$). En la parte inferior se ilustra el proceso para la imagen RGB, haciendo redimensionamiento seguido de una proyección lineal pre-entrenada y congelada ($\mathcal{E}_\omega$). Ambos procesos resultan en embeddings $\mathbf{E}_S$ y $\mathbf{E}_R$ de dimensiones $n \times d$.

ω que incluye al menos la proyección lineal y el *positional encoding*. Por tanto, la colección de *embeddings* RGB se define entonces como $\mathbf{E}_R = \{\hat{\mathbf{e}}_k^\top \in \mathbb{R}^{d_i} \mid k \in \mathbb{N}, 1 \leq k \leq n\}$ con n siendo el número de *tokens*, igual al número de *tokens* espectrales y d la dimensión del espacio de *embedding*, que se mantiene consistente con la dimensión de los *embeddings* espectrales. Debido a que los *tokens* espectrales $\mathbf{E}_S$ y espaciales $\mathbf{E}_R$ deben ser de igual tamaño, es necesario que los bloques vectorizados $\mathbf{b}_k$ sean de mayor tamaño a los parches $\mathbf{p}_k$, es decir $p_h > p_{h'}$ y $p_w > p_{w'}$. A diferencia del *embedding* espectral, los parámetros del *embedding* RGB provienen de un modelo pre-entrenado y se mantienen congelados durante el entrenamiento, el proceso de tokenizacion se ilustra en la Figura 8.

4.2. *Adaptive spectral prompts* (ASP)

Adaptive Spectral Prompts (ASP) se introduce como un mecanismo clave para integrar eficientemente la información espectral en el proceso de segmentación multimodal de materiales. Este enfoque permite adaptar un modelo *transformer* pre-entrenado $\mathcal{M}_{\Theta}$, que comprende h bloques tipo *transformer* $\mathcal{M}_{\Theta^h}$, a la tarea específica de segmentación de materiales y aprovechar la modalidad espectral, manteniendo su robustez incluso en escenarios donde esta modalidad puede estar ausente.

Definimos un conjunto de *prompts* aprendibles[92] $\mathcal{P}$ por cada bloque i como $\mathbf{P}^i = \{\mathbf{p}_l^\top \in \mathbb{R}^{d_i} \mid l \in \mathbb{N}, 1 \leq l \leq l_i\}$, donde $\mathbf{P}^i$ serán los *prompts* que ingresan en el bloque i del *transformer*, l_i es el número de *prompts* aprendibles para el bloque i en específico y l hace referencia a un *prompt* en específico aprendido, d_i es la dimension del *embedding* en el bloque i del *transformer*.

Estos *prompts* $\mathcal{P}$ se inicializan de forma aleatoria y se aprenden durante el entrenamiento, una vez finalizado el entrenamiento quedan fijos en inferencia. Con el fin de que los *prompts* interactúen con la información espectral durante el entrenamiento, y además sea opcional el uso información espectral en inferencia, se construyó el modulo ASP que será usado en cada bloque i del *transformer encoder* $\mathcal{M}_{\theta^{\text{encoder}}}(\cdot)$.

El proceso de integración de los modulos ASP comienza con la concatenación de los *prompts* con los *tokens* espectrales del bloque anterior $\mathbf{E}_S^{i-1}$ (o los *embeddings* espectrales iniciales para el primer bloque $\mathbf{E}_S^0$). Sea $\mathbf{P}^{i-1} \in \mathbb{R}^{l_{i-1} \times d_{i-1}}$ el conjunto de *prompts* de bloque $i-1$ y $\mathbf{E}_s^{i-1} \in \mathbb{R}^{n \times d_{i-1}}$ los *tokens* espectrales del bloque $i-1$, la concatenación entre los *prompts* y los *tokens* será:

$$\mathbf{X}_S^i = [\mathbf{P}^{i-1}; \mathbf{E}_S^{i-1}] \in \mathbb{R}^{(l_{i-1}+n) \times d_{i-1}}. \tag{5}$$

---

[92] Jia et al., ver n. 80.

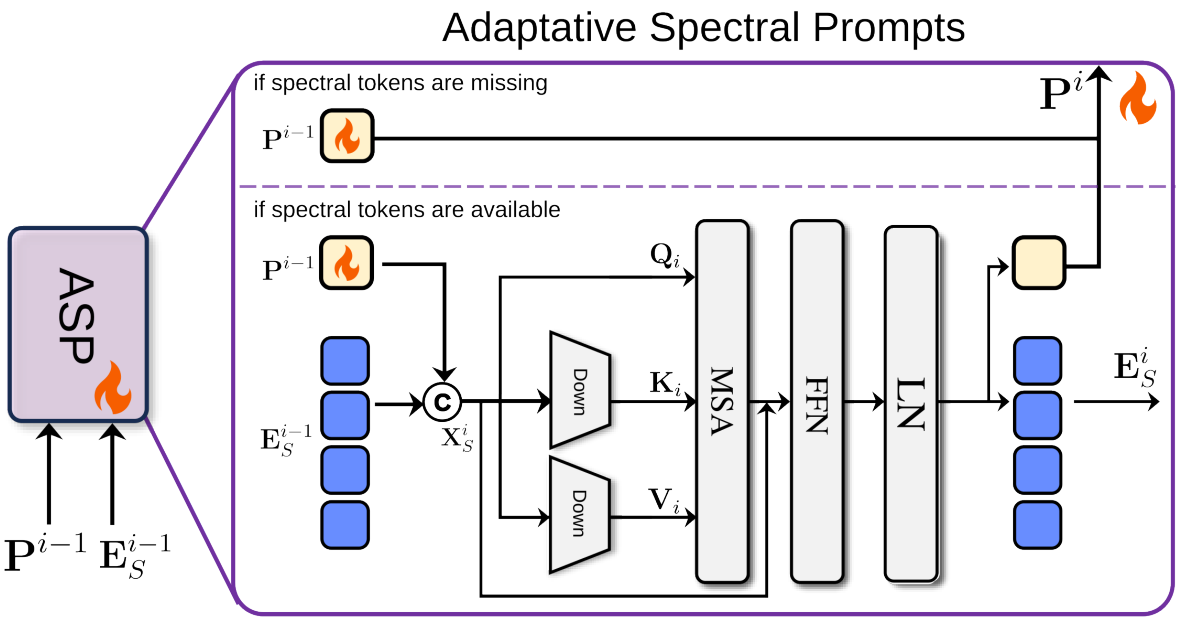

Figura 9. Arquitectura detallada del módulo ASP. El módulo recibe como entrada los *tokens* espectrales $\mathbf{E}_S^{i-1}$ y los prompts $\mathbf{P}^{i-1}$ de la capa anterior. Estos pasan por un módulo de atención multi-cabeza (MSA) con proyecciones reducidas, una capa *feed-forward* (FFN) y una normalización de capa (LN). El módulo produce *prompts* adaptados $\mathbf{P}^i$ y *tokens* espectrales actualizados $\mathbf{E}_S^i$ para la siguiente capa. Los componentes en naranja indican los *prompts*, mientras que los azules representan los *tokens* espectrales.

La representación combinada de *tokens* $\mathbf{X}_s^i$ pasa por el módulo propuesto de ASP en el bloque i, que puede denotarse como:

$$[\hat{\mathbf{P}}^i, \mathbf{E}_S^i] = \text{ASP}^i(\mathbf{X}_S^i) = \text{ASP}^i([\mathbf{P}^{i-1}; \mathbf{E}_S^{i-1}]). \tag{6}$$

El módulo ASP retorna los *prompts* adaptados $\hat{\mathbf{P}}^i$ y los *tokens* espectrales transformados $\mathbf{E}_S^i$. ASP está compuesto por un módulo de atención multi-cabeza (MSA)[93] con reducción espacial[94] con parámetros ψ, seguido de una red *feed-forward* (FFN)

[93] Vaswani et al., ver n. 50.

[94] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". En: *arXiv preprint*

con parámetros δ y normalización por capas (LN)[95]:

$$[\hat{\mathbf{P}}^i, \mathbf{E}_s^i] = \text{LN}^i(\text{FFN}_\delta^i(\text{MSA}_\psi^i(\mathbf{X}_S^i) + \mathbf{X}_S^i)), \tag{7}$$

donde $\text{MSA}_\psi^i(\cdot)$ es el módulo de *multi-head self-attention* en el bloque i definido como:

$$\begin{aligned} \mathcal{A}^i = \text{MSA}_\psi(\mathbf{X}_S^i) &= \text{Concat}(\text{head}^1, \ldots, \text{head}^c)\mathbf{W}^O, \\ \text{con} \quad \text{head}^t &= \text{Attention}(\mathbf{Q}^t, \mathbf{K}^t, \mathbf{V}^t), \\ &= \text{Attention}(\mathbf{X}_S^i \mathbf{W}_Q^t, \mathbf{K}^t, \mathbf{V}^t), \end{aligned} \tag{8}$$

donde $\mathbf{W}_Q^t \in \mathbb{R}^{d_{i-1} \times d_k}$ es la matriz de proyección que se aplica sobre $\mathbf{X}_S^i$ para obtener $\mathbf{Q}^t$, d_{i-1} es la dimensión en el bloque $i-1$ y d_k es la dimensión de proyección para $\mathbf{Q^t}$ y $\mathbf{K^t}$; por tanto, $\mathbf{Q}^t = \mathbf{X}_S^i \mathbf{W}_Q^t \in \mathbb{R}^{n \times d_k}$. La matriz $\mathbf{W}^O \in \mathbb{R}^{(c \times d_v) \times d_i}$ proyecta a las dimensiones originales d_i del bloque del *transformer*, donde d_v es de la dimensión de proyección para $\mathbf{V}^t$ y c es el numero de cabezas *(heads)*, cada una de las cabezas de atención head^t se calcula mediante la operación de *Attention*, que utiliza las proyecciones de consultas $\mathbf{Q}$, claves $\mathbf{K}$ y valores $\mathbf{V}$, de la siguiente manera:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{9}$$

Este proceso de *Attention* tiene una complejidad computacional $O(n^2)$, lo que puede resultar prohibitivo para el procesamiento de imágenes de gran tamaño, por esta

---

arXiv:2106.09685 (2021).

[95] Jimmy Lei Ba. “Layer normalization”. En: *arXiv preprint arXiv:1607.06450* (2016).

razón, proponemos aplicar un proceso de reducción de secuencia[96,97] para $\mathbf{K}$ y $\mathbf{V}$. Este proceso usa un factor de reducción r para reducir el tamaño n de la secuencia así:

$$\hat{\mathbf{K}}^t = \text{Reshape}\left(\frac{n}{r}, d_{i-1} \cdot r\right)(\mathbf{X}_S^i), \tag{10}$$

$$\mathbf{K}^t = \hat{\mathbf{K}}^t \mathbf{W}_K^t, \tag{11}$$

donde Reshape$(\frac{n}{r}, d_{i-1} \cdot r)(\cdot)$ es la operación que se refiere a redimensionar $(\cdot)$ para asi obtener $\hat{K^t} \in \mathbb{R}^{\frac{n}{r} \times (d_{i-1} \cdot r)}$, para luego pasar por la proyección lineal $W_K^t \in \mathbb{R}^{(d_{i-1} \cdot r) \times d_k}$, obteniendo $K^t \in \mathbb{R}^{\frac{n}{r} \times d_k}$. El proceso es el mismo para el *value* V^t, haciendo Reshape y proyectando con la matriz $W_V^t \in \mathbb{R}^{\frac{n}{r} \times d_v}$, donde d_v es la dimensión de proyección de $\mathbf{V^t}$. Como resultado, la complejidad del mecanismo de *Attention* definido en la Ecuación 9 es reducida de $O(n^2)$ a $O(\frac{n^2}{r})$.

Este mecanismo propuesto en[98] permite capturar información relevante desde distintos subespacios a través de múltiples cabezas h, mejorando la capacidad del modelo para extraer características contextuales importantes de manera eficiente, mientras que la reducción dimensional permite mantener el costo computacional bajo de la operación. Finalmente, aplicamos una FFN con parámetros ψ que está definida como:

$$\mathcal{Z}^i = \text{FFN}_\psi^i(\mathcal{A}^i) = \text{MLP}^i(\text{GELU}^i(\text{MLP}^i(\mathcal{A}^i))) + \mathcal{A}^i, \tag{12}$$

donde $\mathcal{A}^i$ son las características extraídas a partir del modulo MSA en la Ecuación

---

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

97 Hu et al., ver n. 94.

98 Vaswani et al., ver n. 50.

8, MLP es un perceptron multicapa que mantiene las mismas dimensiones de la entrada y GELU es una función de activación.[99] Seguido de aplicar la FFN se aplica una normalización por capas, dada por:

$$[\hat{\mathbf{P}}^i, \mathbf{E}_s^i] = \mathsf{LN}^i(\mathcal{Z}^i), \tag{13}$$

donde $\hat{\mathbf{P}}^i$ son los *prompts* $\mathbf{P}^{i-1}$ que ingresaron al modulo ASP adaptados con la información spectral; así mismo, $\mathbf{E}_S^i$ serán los *tokens* espectrales transformados que se guardan para el siguiente bloque del *transformer* $i+1$.
De forma simplificada, podemos representar todo este proceso como:

$$[\hat{\mathbf{P}}^i, \mathbf{E}_S^i] = \mathsf{ASP}_i([\mathbf{P}^{i-1}; \mathbf{E}_S^{i-1}]), \tag{14}$$

donde ASP engloba todas las operaciones descritas anteriormente, e i es el bloque actual del *transformer*. De estas representaciones transformadas, se toma $\mathbf{P}^i$ que son los *tokens* correspondientes a los *prompts* adaptados. Estos *prompts* adaptados $\mathbf{P}^i$ se integran con los tokens RGB $\mathbf{E}_R^{i-1} \in \mathbb{R}^{n \times d}$ de la capa anterior mediante concatenación:

$$\mathbf{X}_R^i = [\hat{\mathbf{P}}^i; \mathbf{E}_R^{i-1}] \in \mathbb{R}^{(l+n) \times d}. \tag{15}$$

La representación combinada $\mathbf{X}_R^i$ se procesa a través del i-ésimo bloque $\mathcal{M}_{\Theta^i}$ del *transformer* pre-entrenado y congelado $\mathcal{M}_\Theta$:

$$[\_, \mathbf{E}_R^{i+1}] = \mathcal{M}_{\Theta^i}(\mathbf{X}_R^i), \tag{16}$$

los *prompts* que se ingresaron contenados mediante $\mathbf{X}_R^i$ son descartados en la

---

[99] Dan Hendrycks y Kevin Gimpel. “Gaussian error linear units (gelus)”. En: *arXiv preprint arXiv:1606.08415* (2016).

salida del bloque $\mathcal{M}_{\Theta^i}$. Este proceso se repite para cada bloque del *transformer encoder-decoder* $\mathcal{M}_\Theta$, utilizando los prompts adaptados del bloque anterior $\mathbf{P}^{i-1}$. Es importante destacar que los bloques del *transformer* $\mathcal{M}_{\Theta^i}$ se mantienen congelados durante todo el proceso, lo que permite una adaptación eficiente del modelo pre-entrenado sin necesidad de ajustar sus parámetros internos y reduciendo la cantidad de parámetros a entrenar.

En escenarios donde la modalidad espectral está ausente, el ASP demuestra su flexibilidad. En estos casos de ausencia de modalidad espectral el ASP da como salida los mismos *prompts* $\mathbf{P}^{i-1}$ que se le ingresan sin ningún computo adicional:

$$[\hat{\mathbf{P}}^i, \_] = \text{ASP}^i([\mathbf{P}^{i-1}, \_]). \tag{17}$$

En estos casos, los *prompts* aprendibles $\mathbf{P}^i = \hat{\mathbf{P}}^i$ en el bloque i se concatenan directamente con los *tokens* RGB:

$$\mathbf{X}_R^i = [\mathbf{P}^i; \mathbf{E}_R^{i-1}] \in \mathbb{R}^{(l+n)\times d}. \tag{18}$$

Esta representación se procesa a través del bloque del *transformer* de la misma manera que en el caso multimodal, permitiendo que el modelo mantenga su funcionalidad incluso sin información espectral. El enfoque ASP proporciona una solución elegante para adaptar modelos pre-entrenados a tareas de segmentación multimodal de materiales. Al optimizar únicamente los parámetros de los *prompts* y los módulos ASP durante el entrenamiento, se logra una adaptación eficiente y efectiva, manteniendo la capacidad del modelo para manejar tanto escenarios multimodales como unimodales con una cantidad mínima de parámetros ajustables.

4.3. Entrenamiento con modalidad faltante

Para abordar el desafío de la segmentación multimodal de materiales en escenarios donde una modalidad puede estar ausente, se implementó una estrategia de entrenamiento robusta que incorpora el concepto de *modality dropout* inspirado en.[100] Este enfoque permite que el modelo sea capaz de manejar situaciones de modalidad faltante durante la inferencia, mostrándole el comportamiento de modalidad faltante durante el entrenamiento, para que así sea robusto en la inferencia. El marco de trabajo utiliza un conjunto de datos $\mathcal{D} = (\hat{\mathcal{I}}_S^i, \mathcal{I}_R^i, \mathbf{Y}^i)_{i=1}^N$, con N muestras, donde $\mathbf{Y}^i \in \{1, ..., c\}^{H \times W}$ es la máscara de segmentación correspondiente con c clases de materiales. Durante el entrenamiento, se aplica el *modality dropout* con una probabilidad p_d. Cuando se activa, se omite completamente la modalidad espectral $\mathcal{I}_S$, forzando al modelo a realizar la segmentación basándose únicamente en la información RGB. Este proceso se puede describir matemáticamente como:

$$\mathcal{I}_S = \begin{cases} \hat{\mathcal{I}}_S, & \text{con probabilidad } 1 - p_d \\ \emptyset, & \text{con probabilidad } p_d \end{cases} \tag{19}$$

donde $\hat{\mathcal{I}}_S$ es el cubo espectral y $\emptyset$ es conjunto vació, $\mathcal{I}_S$ es la entrada espectral utilizada durante el entrenamiento, que puede ser $\hat{\mathcal{I}}_S$ o $\emptyset$ con probabilidad p_d.
Es importante destacar que todo el *encoder* $\mathcal{M}_{\Theta^{\text{encoder}}}$ del *transformer* pre-entrenado $\mathcal{M}_\Theta$ se mantiene congelado durante el proceso de entrenamiento. Esto incluye los bloques de MSA y las capas FFN del *encoder*. Sin embargo, el *decoder* del *transformer* $\mathcal{M}_{\Theta^{\text{decoder}}}$, sí se entrena junto con los módulos ASP$(\cdot)$ y los *prompts* espectrales. Esta estrategia permite una adaptación eficiente del modelo $\mathcal{M}_\Theta$ a la tarea de seg-

---

[100] Yi-Lun Lee et al. "Multimodal prompting with missing modalities for visual recognition". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 14943-14952.

mentación multimodal de materiales, llevando a aprender representaciones robustas que son efectivas tanto en escenarios multimodales como en aquellos donde solo está disponible la información RGB, además de aprovechar el conocimiento previo del modelo pre-entrenado.

Al alternar entre la presencia y ausencia de la modalidad espectral durante el entrenamiento, el modelo desarrolla la capacidad de adaptarse dinámicamente a la información disponible, mejorando su generalización y rendimiento en situaciones de modalidad faltante durante la inferencia.

Por ultimo, la codificación completa del algoritmo propuesto puede ser consultada en el siguiente enlace: `https://github.com/Factral/spectral-transformer`

5. RESULTADOS

5.1. Base de datos

La evaluación de métodos de segmentación de materiales en escenas naturales que integren información espectral y RGB se ve frecuentemente limitada por la escasez de conjuntos de datos que proporcionen de manera simultánea imágenes RGB, datos espectrales y mapas de segmentación de materiales en forma emparejada.[101] Esta carencia representa un desafío significativo. Ante este panorama, en este trabajo elegimos el conjunto de datos *Light Industrial Building HSI (LIB-HSI)*[102] como el conjunto de datos principal para todas nuestras pruebas y evaluaciones. LIB-HSI ofrece la combinación requerida de datos, proporcionando imágenes de fachadas en escenas naturales. LIB-HSI se compone de un conjunto $\mathcal{D} = (\mathcal{I}_S^i, \mathcal{I}_R^i, \mathbf{Y}^i)_{i=1}^N$ que comprende un total de $N = 513$ imágenes. Cada cubo de datos hiperspectral $\mathcal{I}_S^i$ tiene una resolución espacial de 512x512 píxeles y 204 bandas, abarcando un rango espectral de 400 a 1000 nm. Esta riqueza espectral permite una caracterización precisa de los materiales presentes en las escenas. El conjunto de datos está dividido en tres sub-conjuntos: *train* que contiene 392 imágenes, *validation* con 45 imágenes y test con 75 imágenes.

Para el presente trabajo, se hizo un muestreo de 64 bandas del rango de 400 a 900nm. Este rango fue seleccionado para reducir la complejidad computacional sin sacrificar significativamente la información relevante para la segmentación de materiales. A través de un proceso de media móvil, se redujo el número de bandas espectrales de

---

[101] Yuwen Heng et al. “MatSpectNet: Material Segmentation Network with Domain-Aware and Physically-Constrained Hyperspectral Reconstruction”. En: *arXiv preprint arXiv:2307.11466* (2023).

[102] Habili et al., ver n. 37.

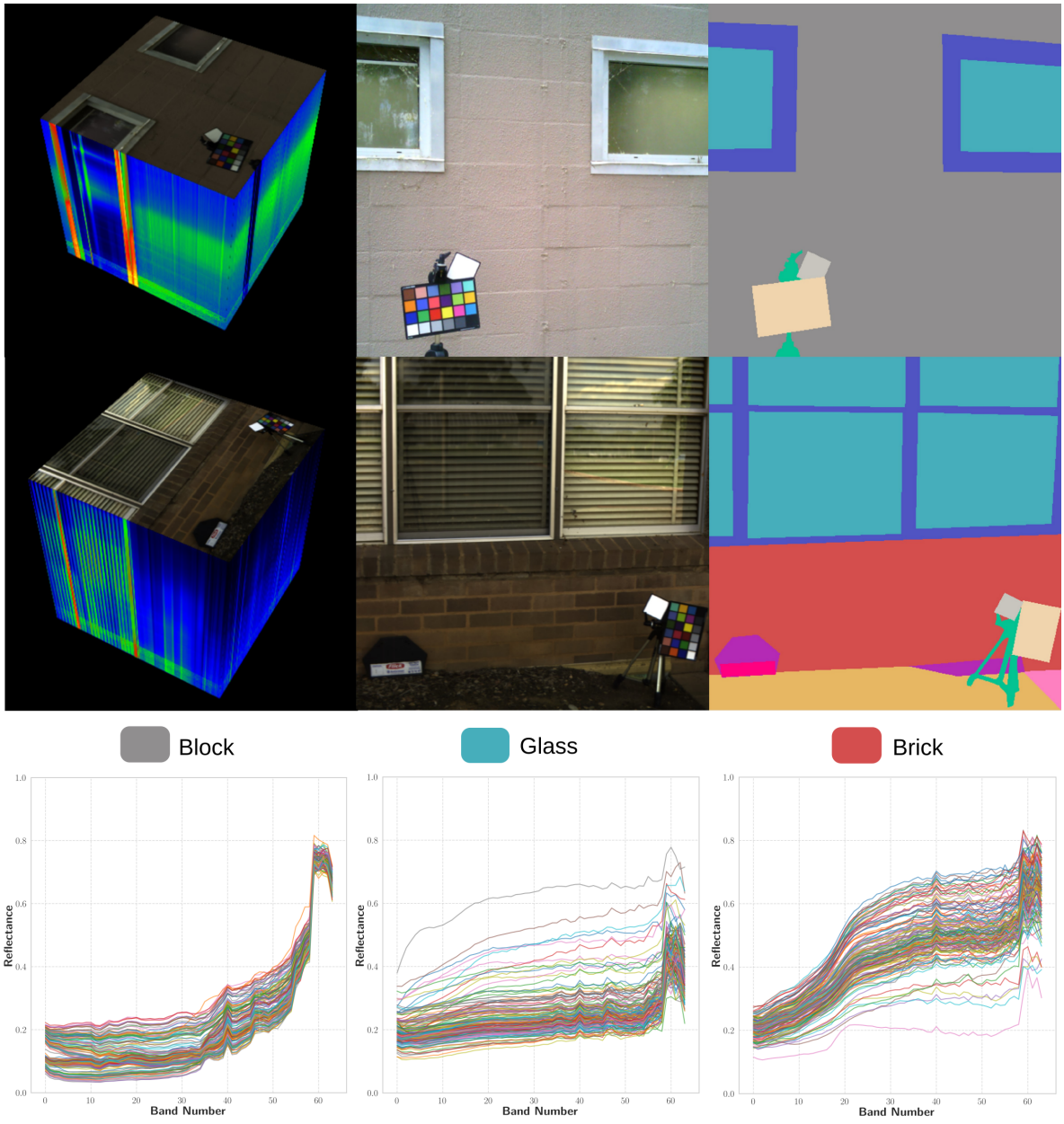

Figura 10. Ejemplos de imágenes del dataset LIB-HSI. Las dos primeras filas muestran, de izquierda a derecha, las imágenes RGB, las imágenes hiperespectrales correspondientes y los mapas de segmentación de materiales. En la última fila se presentan ejemplos de firmas espectrales para materiales seleccionados como 'Block', 'Glass' y 'Brick', obtenidas a partir de 150 píxeles seleccionados aleatoriamente de cada material identificado en la imagen del dataset. Las firmas espectrales incluyen 64 bandas, calculadas con la media móvil, que cubren el rango de 400 nm a 1000 nm. Adaptado de.[103]

204 a 64, manteniendo una representación espectral suficientemente detallada para la discriminación de materiales. El dataset LIB-HSI se distingue por su exhaustiva categorización de materiales, con un total de 44 clases diferentes. Estas clases abarcan una amplia gama de materiales comúnmente encontrados en entornos industriales y urbanos, incluyendo varios tipos de metales, concreto, madera, vidrio, plásticos, y vegetación. La diversidad de clases presenta un desafío significativo para los algoritmos de segmentación, especialmente considerando la variabilidad en la apariencia de estos materiales bajo diferentes condiciones de iluminación y ángulos de visión.

Antes de usar el conjunto de datos, se realizó un análisis detallado de la distribución de clases en los conjuntos de entrenamiento, validación y *test* originales, este análisis reveló desequilibrios significativos que podrían afectar el rendimiento y la generalización del modelo. En particular, se contabilizó la presencia y frecuencia de cada clase en los diferentes conjuntos. Los resultados obtenidos mostraron que dos clases en particular estaban presentes en solo uno ó dos de los subconjuntos del conjunto de datos. En el Cuadro 1 se muestran los resultados obtenidos para las clases que no están presentes en todos los subconjuntos del conjunto de datos

		Número de pixeles		
Clase	**Imágenes**	**Train**	**Validation**	**Test**
Wood Ground	1	116257	0	0
Door-plastic	2	177131	0	0

Cuadro 1. Distribución de clases mal representadas en el dataset LIB-HSI. La columna 'Clase' indica el material, 'Imágenes' muestra el número total de imágenes que contienen la clase, y las columnas 'Train', 'Validation' y 'Test' representan el número de píxeles de cada clase en los respectivos conjuntos.

Encontramos que la clase *Wood Ground* y *Door-plastic* estaban mal representadas en el dataset. Especificamente, la clase *Wood Ground* estaba presente en solo una imagen de toda la base de datos y se encontraba solo en el subconjunto de *train* con 116257 pixeles. Asi mismo, la clase *Door-plastic* se encontraba en 2 imágenes

diferentes con 177131 pixeles. La presencia de clases exclusivamente en el conjunto de entrenamiento puede llevar a un sobreajuste del modelo, comprometiendo su capacidad de generalización. Este fenómeno es particularmente problemático en el contexto de la segmentación de materiales, donde la diversidad de apariencias y condiciones de iluminación ya supone un reto considerable.[104] El desequilibrio afecta directamente las métricas de evaluación estándar en tareas de segmentación semántica, precisión media (del inglés *mean accuracy* mACC) o el *Intersection over Union* (IoU) por clase.[105] Las métricas globales podrían enmascarar el rendimiento deficiente en clases infrarrepresentadas, mientras que las métricas por clase podrían mostrar resultados engañosamente optimistas para estas categorías en el conjunto de entrenamiento, siendo inaplicables en los conjuntos de validación y prueba donde estas clases están ausentes.

5.1.1. *Split* propuesto

Debido a que la tarea que se está abordando es segmentación densa de materiales en una escena, no se puede hacer una re-distribución de pixeles por separado, en todo caso se puede hacer una re-distribución de imágenes. Para esto, se propuso un nuevo *split* del dataset entero. Las modificaciones principales incluyeron la eliminación de las clases *Wood Ground* y *Door-plastic* debido a su mala representación y presencia en un solo subconjunto, esto es, eliminar las 3 imágenes que contenían estas clases, pasando de tener 513 imágenes a 510 imágenes en total.
El nuevo *split* se diseñó con el objetivo de mantener una distribución de clases lo más similar posible entre los tres conjuntos, ademas de mantener el mismo número de

---

[104] Bell et al., ver n. 33.

[105] Gabriela Csurka et al. “What is a good evaluation measure for semantic segmentation?.” En: *Bmvc*. Vol. 27. 2013. Bristol. 2013, págs. 10-5244.

imágenes en los subconjuntos de *validation* y *test* que el conjunto original. El proceso de redistribución se llevó a cabo mediante un algoritmo iterativo que analizaba la composición de clases en cada imagen y realizaba intercambios entre los conjuntos para optimizar el equilibrio global minimizando la *KL divergence*[106] primero entre *train* y *validation*, y luego entre *train* y *test*. La nueva distribución resultante asegura que todas las clases restantes (42 en total) estén representadas en los tres conjuntos: *train*, *validation* y *test*. Esto permite una evaluación más justa y robusta del modelo de segmentación de materiales, reduciendo el riesgo de sobreajuste a clases específicas y mejorando la capacidad del modelo para generalizar a nuevas imágenes.

5.2. Métricas de evaluación

Para evaluar el rendimiento de nuestro método propuesto, adoptamos las métricas estándar ampliamente utilizadas en el estado del arte de la segmentación semántica y de materiales. Específicamente, empleamos dos métricas principales: la precisión absoluta *(absolute accuracy)* y la media del *Intersection over Union (mean IoU)* sobre las clases.

Precisión Absoluta: Se define como la proporción de píxeles correctamente clasificados en relación con el total de píxeles en la imagen. Se calcula como:

$$\text{Acc} = \frac{\sum_{c=1}^{C}(TP)_c}{M}, \tag{20}$$

donde $(TP)_c$ denota el número de píxeles correctamente clasificados para la clase c, C es el número total de clases y M es el número total de píxeles en la imagen. Esta fórmula refleja la precisión general del modelo, ya que contabiliza todos los píxeles correctamente clasificados sin importar su clase específica, proporcionando

---

[106] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

una evaluación directa del rendimiento del modelo sobre la totalidad de la imagen.

Intersection over Union:[107] Se utiliza para evaluar la precisión de la segmentación por clase. El IoU para una clase específica se define como la intersección entre los píxeles predichos y los píxeles verdaderos, dividida por la unión de ambos. Se formula como

$$\text{IoU}_c = \frac{(TP)_c}{(TP)_c + (FP)_c + (FN)_c}, \tag{21}$$

donde c representa la clase específica. La métrica de mean IoU se calcula promediando el IoU sobre todas las clases, de la siguiente manera:

$$\text{mean IoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c, \tag{22}$$

donde C es el número total de clases. Estas métricas no solo permiten evaluar la efectividad del modelo de segmentación, sino que también son fundamentales para la comparación con otros métodos en la literatura, garantizando una interpretación adecuada de los resultados obtenidos.

La elección de estas métricas nos permite evaluar tanto el rendimiento global del modelo como su capacidad para segmentar correctamente cada clase de material individual. Esto es crucial en nuestro contexto, dado el desequilibrio de clases observado en el dataset LIB-HSI y la importancia de identificar correctamente incluso los materiales menos frecuentes en las escenas industriales.

5.3. Simulaciones

5.3.1. Estudios de ablación

Para evaluar rigurosamente el rendimiento de nuestro método propuesto y validar

---

[107] Csurka et al., ver n. 105.

la hipótesis de que la información espectral incorporada en la arquitectura mejora significativamente la segmentación de materiales, realizamos un exhaustivo estudio de ablación. Este estudio compara nuestro enfoque con otros métodos alternativos, proporcionando una visión profunda de la contribución de cada componente de nuestra arquitectura.

Como base para nuestros experimentos, seleccionamos la arquitectura *SegFormer*,[108] específicamente el modelo *SegFormer-B3*. Esta elección se justifica por su rendimiento sobresaliente en tareas de segmentación semántica y su balance entre capacidad y eficiencia computacional, lo que nos permite realizar un entrenamiento exhaustivo con los recursos disponibles. *SegFormer-B3* tiene 4 bloques $h = 4$ de *transformer* en su *encoder* $\mathcal{M}_{\theta^{encoder}}$ con dimensiones $d_i = [64, 128, 320, 512]$, así mismo se dejo $d_v = d_i$, ademas se dejó fijo el factor de reducción $r = [64, 16, 4, 1]$ desde el bloque 1 al bloque 4, además dejamos fijo el numero de *prompts* aprendibles $\mathbf{P}^i$ en 10 para cada bloque i del modelo. Para el proceso de *embeddings* para el cubo espectral, fijamos $p_h = 8$, $p_w = 8$ y $p_b = 16$, para los *embeddings* de la imagen RGB, se dejo fijo $p'_h = 4$ y $p'_w = 4$ Todos los experimentos se llevaron a cabo en una GPU NVIDIA RTX 3090, utilizando un tamaño de lote *(batch size)* de 16. Para optimizar el uso de memoria y acelerar el entrenamiento, empleamos la técnica de precisión mixta de 16 bits para los pesos del modelo.[109] El entrenamiento se realizó utilizando el optimizador *AdamW*[110] con una tasa de aprendizaje inicial de 1e-4, junto con un programador de tasa de aprendizaje de tipo coseno *(cosine scheduler)* para un ajuste dinámico durante el entrenamiento.

---

[108] Xie et al., ver n. 14.

[109] Paulius Micikevicius et al. “Mixed Precision Training”. En: *International Conference on Learning Representations*. 2018.

[110] I Loshchilov. “Decoupled weight decay regularization”. En: *arXiv preprint arXiv:1711.05101* (2017).

Es importante destacar que todos los experimentos de ablación se realizaron sobre nuestra versión modificada del dataset LIB-HSI, denominada LIB-HSI-fixed, que aborda los problemas de desequilibrio de clases identificados previamente.

Realizamos cuatro experimentos principales para evaluar diferentes aspectos de nuestra propuesta:

Full Fine-tuning: En este experimento, realizamos un ajuste fino completo de la arquitectura SegFormer-B3, partiendo de pesos pre-entrenados en ImageNet. Este enfoque representa la línea base de comparación, permitiendo que todos los parámetros del modelo se ajusten a la tarea de segmentación de materiales.

Prompt Tuning RGB: Este experimento implementa la técnica de *prompt tuning*[111] utilizando únicamente información RGB. Aquí, introducimos los *prompts* aprendibles que se concatenan con los *embeddings* de entrada de cada bloque, mientras mantenemos congelados los pesos del modelo base. Este enfoque evalúa la eficacia de las técnicas de *prompt tuning* en el contexto de la segmentación de materiales sin información espectral adicional.

Prompt Tuning Spectral sin *modality dropout*: Evaluamos nuestro método propuesto con el modulo *ASP* junto con la información espectral sin el uso de la técnica del *modality dropout*.

Prompt Tuning Spectral (Nuestro Método): Finalmente, evaluamos nuestro método propuesto, que incorpora el módulo *ASP* junto con la información espectral. Este experimento incluye el uso de *modality dropout* con $p_d = 0{,}2$ durante el entrenamiento para mejorar la robustez del modelo ante la ausencia de información espectral en inferencia.

Los resultados de nuestros experimentos de ablación se presentan en el Cuadro 2. Analizando los resultados, observamos tendencias significativas:. El metodo de *Full*

---

[111] Lester, Al-Rfou y Constant, ver n. 82.

Método	Acc	Promedio por clase IoU	Parametros
Full Fine-tuning	85,94	46,98	44,7 millones
Prompt-tuning RGB	86,4	54,92	3,3 millones
Nuestro *(sin modality dropout)*	88,16	54,95	11 millones
Nuestro *(Prompt Tuning Spectral)*	**88,54**	**56,84**	11 millones

Cuadro 2. Resultados del estudio de ablación. Se comparan tres configuraciones: *Fine-tuning* completo, *Prompt-tuning* solo con RGB, nuestro método sin *modality dropout*, y nuestro método propuesto *(Prompt Tuning Spectral)* junto a *modality dropout*. Se muestran la precisión promedio *(Accuracy)* y el IoU promedio por clase *(Average Class IoU)* para cada configuración.

Fine-tuning logra una precisión (Accuracy) del 85.94 % y un IoU promedio por clase de 46.98 %. A pesar de permitir la actualización de todos los parámetros del modelo, su rendimiento es inferior a los métodos basados en *prompt tuning*. Esto sugiere que, para la tarea específica de segmentación de materiales, la adaptación completa del modelo puede llevar a un sobreajuste, especialmente considerando el tamaño limitado del conjunto de datos LIB-HSI-fixed. El metodo de *Prompt Tuning solo con información RGB* muestra una mejora significativa, alcanzando una precisión del 86.4 % y un IoU promedio por clase de 54.92 %. Este incremento sustancial en el rendimiento, utilizando solo 0.4 millones de parámetros entrenables, demuestra la eficacia de las técnicas de *prompt tuning* para la segmentación de materiales. La mejora en el IoU promedio sugiere que este enfoque es particularmente efectivo para manejar el desequilibrio de clases presente en el conjunto de datos.

Finalmente, nuestro método propuesto *Prompt Tuning Spectral* que incorpora información espectral a través del módulo ASP junto a la técnica de *modality dropout*, logra el mejor rendimiento con una precisión del 88.54 % y un IoU promedio por clase de 56.84 %. Este resultado superior se obtiene con solo 0.6 millones de parámetros entrenables, demostrando la eficiencia y efectividad de nuestro método. La mejora sobre el *Prompt Tuning* RGB (+2.14 % en precisión y +1.92 % en IoU) valida nuestra hipótesis de que la incorporación de información espectral, incluso de

manera embebida, puede mejorar significativamente la segmentación de materiales. Es notable que nuestro método logre el mejor rendimiento con un aumento no tan amplio en cuanto a parametros comparado con el Prompt Tuning RGB. Esto sugiere que la arquitectura ASP propuesta es altamente eficiente en su uso de la información espectral, permitiendo una mejora significativa en el rendimiento sin un aumento sustancial en la complejidad del modelo.

La superioridad de los enfoques basados en *prompt tuning* sobre el *fine-tuning* completo resalta la importancia de estrategias de adaptación eficientes para tareas específicas como la segmentación de materiales. Estos resultados sugieren que, para conjuntos de datos de tamaño moderado como LIB-HSI-fixed, las técnicas de *prompt tuning* pueden ser más efectivas que la adaptación completa del modelo, posiblemente debido a su capacidad para evitar el sobreajuste mientras se enfocan en aprender representaciones específicas de la tarea.

Tras evaluar el rendimiento general de nuestro método, procedimos a investigar su comportamiento bajo diferentes condiciones de entrada durante la fase de inferencia. Específicamente, comparamos el rendimiento del modelo cuando se le proporciona únicamente información RGB frente a cuando se le suministra tanto información RGB como espectral. Este análisis es crucial para evaluar la robustez y adaptabilidad de nuestro método en escenarios donde la información espectral podría no estar disponible durante la inferencia. El cuadro 3 presenta los resultados de estas pruebas:

		Promedio por clase
Entrada	**Acc**	**IoU**
RGB	$88{,}54$	$56{,}84$
RGB + Espectral	$\mathbf{88{,}63}$	$\mathbf{56{,}95}$

Cuadro 3. Comparación del rendimiento del modelo en inferencia con y sin información espectral. Se muestran los resultados de precisión *(Accuracy)* e IoU promedio por clase *(Average Class IoU)* para dos configuraciones de entrada: solo RGB (modalidad faltante) y RGB + Espectral (modalidad completa).

Los resultados muestran un rendimiento notablemente consistente entre ambas con-

figuraciones de entrada. Cuando se utiliza únicamente información RGB durante la inferencia, el modelo alcanza una precisión del $88{,}54\,\%$ y un IoU promedio por clase de $56{,}84\,\%$. Por otro lado, al incorporar tanto información RGB como espectral, se observa una ligera mejora, con una precisión del $88{,}63\,\%$ y un IoU promedio por clase de $56{,}95\,\%$. Esta diferencia marginal en el rendimiento ($0{,}09\,\%$ en precisión y $0{,}11\,\%$ en IoU) entre las dos configuraciones de entrada es particularmente interesante. Por un lado, demuestra la robustez de nuestro método, capaz de mantener un alto nivel de rendimiento incluso cuando se le priva de la información espectral durante la inferencia. Esto es un testimonio de la efectividad de nuestra estrategia de *modality dropout* durante el entrenamiento, que ha permitido al modelo aprender representaciones robustas que no dependen exclusivamente de la información espectral. Por otro lado, la mejora, aunque ligera, observada al incluir información espectral durante la inferencia, valida nuestra hipótesis inicial sobre el valor de incorporar datos espectrales en la tarea de segmentación de materiales. Este incremento en el rendimiento sugiere que el modelo es capaz de utilizar efectivamente la información espectral adicional para refinar sus predicciones, especialmente en casos ambiguos o desafiantes. La consistencia en el rendimiento entre ambas configuraciones tiene implicaciones prácticas significativas. Sugiere que nuestro modelo podría desplegarse eficazmente en una variedad de escenarios, desde aplicaciones donde solo está disponible la información RGB, hasta entornos más especializados donde se puede acceder a datos espectrales. Esta flexibilidad es particularmente valiosa en contextos del mundo real, donde la disponibilidad de sensores espectrales puede variar.

5.3.2. Resultados cuantitativos

Para contextualizar el rendimiento de nuestro método propuesto dentro del panorama actual de la segmentación de materiales, realizamos una comparación exhaustiva con los métodos del estado del arte. Es importante señalar que, si bien nuestro enfoque se desarrolló y evaluó inicialmente en el conjunto de datos LIB-HSI-fixed, la

mayoría de los métodos existentes en la literatura utilizan el dataset LIB-HSI original. Para garantizar una comparación justa y directa, entrenamos una versión adicional de nuestro modelo utilizando el *split* estándar del LIB-HSI.

		Promedio por clase
Método	**Acc**	**IoU**
FCN[112]	$82,9$	$44,3$
SFM+HRnet[113]	$86,47$	$48,37$
CSSF+DeepLabV3[114]	$---$	$51,2$
Ours (Prompt Tuning Spectral)	$\mathbf{88,36}$	$\mathbf{53,28}$

Cuadro 4. Comparación de rendimiento entre diferentes métodos de segmentación de materiales para el dataset LIB-HSI. Se muestran los resultados de precisión promedio *(Average Accuracy)* y IoU promedio por clase *(Average Class IoU)* para cada método evaluado en el conjunto de datos LIB-HSI.

El Cuadro 4 presenta una comparación detallada de nuestro método con los enfoques del estado del arte. Los resultados demuestran claramente la superioridad de nuestro método propuesto en comparación con los enfoques existentes. Nuestro modelo supera consistentemente a todas las técnicas anteriores tanto en precisión global como en IoU promedio por clase, estableciendo un nuevo estado del arte en la segmentación de materiales en el dataset LIB-HSI. Específicamente, nuestro método logra una mejora del $1{,}89\,\%$ en precisión y del $2{,}08\,\%$ en IoU promedio por clase sobre el mejor método anterior. Es notable que nuestro método logre este rendimiento superior manteniendo una eficiencia computacional comparable o incluso mejor que muchos de los métodos anteriores. Esto subraya no solo la efectividad de

---

[112] Nariman Habili et al. “A hyperspectral and RGB dataset for building façade segmentation”. En: *European Conference on Computer Vision*. Springer. 2022, págs. 258-267.

[113] Fabian Perez y Hoover Rueda-Chacón. “Beyond Appearances: Material Segmentation with Embedded Spectral Information from RGB-D imagery”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2024, págs. 293-301.

[114] Zhuoran Du et al. “Exploring the applicability of spectral recovery in semantic segmentation of RGB images”. En: *IEEE Transactions on Multimedia* (2024).

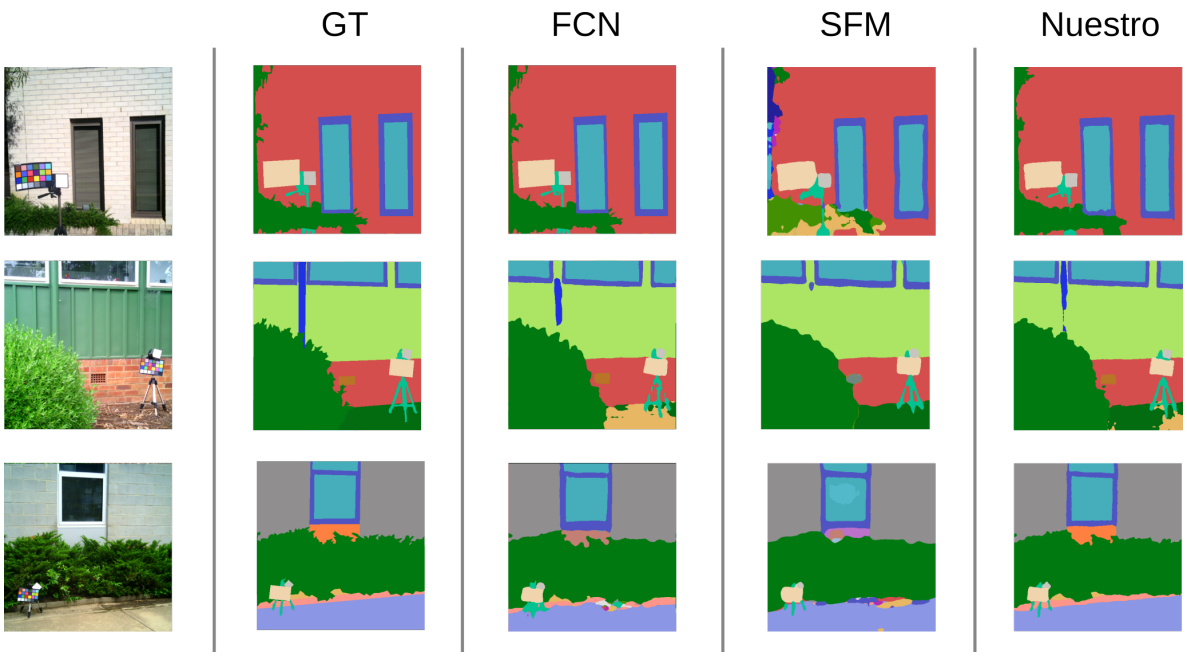

Figura 11. Comparación cualitativa de los resultados de segmentación de materiales. De izquierda a derecha: imagen de entrada RGB, ground truth (GT), resultados de FCN,[115] SFM+HRnet [108], y nuestro método propuesto. Para el método CSSF+DeepLabV3 [109] no se muestran resultados visuales debido a la falta de acceso al modelo y a que sus resultados visuales publicados corresponden a imágenes diferentes. Nuestro método demuestra una segmentación más precisa y coherente en comparación con los otros enfoques.

nuestro enfoque en términos de precisión, sino también su viabilidad práctica para aplicaciones en el mundo real.

5.3.3. Resultados cualitativos

Para complementar nuestro análisis cuantitativo, presentamos una evaluación cualitativa de los resultados de segmentación obtenidos por nuestro método en comparación con otros enfoques del estado del arte. La Figura 11 muestra ejemplos visuales de segmentación de materiales en imágenes del conjunto de datos LIB-HSI.

---

[115] Nariman Habili et al. “A hyperspectral and RGB dataset for building façade segmentation”. En: *European Conference on Computer Vision*. Springer. 2022, págs. 258-267.

Los resultados cualitativos demuestran la eficacia de nuestro método en la segmentación precisa de materiales en escenas complejas. En comparación con los otros enfoques, nuestro método muestra una mayor coherencia en la segmentación, especialmente en áreas con transiciones sutiles entre materiales y en regiones con texturas complejas. Se observa una mejora notable en la delimitación de bordes entre diferentes materiales y en la correcta clasificación de áreas pequeñas o detalles finos. En particular, nuestro método demuestra un rendimiento superior en la segmentación de elementos arquitectónicos como ventanas, puertas y detalles estructurales, donde los otros métodos tienden a producir segmentaciones más fragmentadas o imprecisas. Además, se aprecia una mejor discriminación entre materiales similares, como diferentes tipos de metales o superficies reflectivas, lo que sugiere que nuestro enfoque aprovecha eficazmente la información espectral incorporada.

5.3.4. Resultados experimentales cualitativos

Para evaluar la capacidad de generalización de nuestro modelo en escenarios del mundo real, se realizó una validación cualitativa utilizando imágenes capturadas en el campus universitario, fuera del conjunto de datos de entrenamiento. Esta prueba tiene como objetivo demostrar la robustez y aplicabilidad del método propuesto en entornos no controlados. Se capturaron dos imágenes representativas de escenas arquitectónicas comunes en el entorno universitario utilizando la cámara de un Ipad Pro, cada imagen fue redimensionada a 512×512. Estas imágenes fueron procesadas por nuestro modelo entrenado, sin ningún ajuste adicional o *fine-tuning* específico para estas nuevas escenas.

Como se puede observar en la Figura 12, nuestro modelo demuestra una notable capacidad para generalizar a escenas no vistas previamente. La segmentación resultante muestra una correcta identificación y delimitación de los principales materiales presentes en las imágenes, como ladrillo, concreto, vegetación y metal, entre otros. Es importante destacar que esta evaluación es puramente cualitativa, ya que no se

dispone de etiquetas (*ground truth*) para estas imágenes capturadas en el campus.

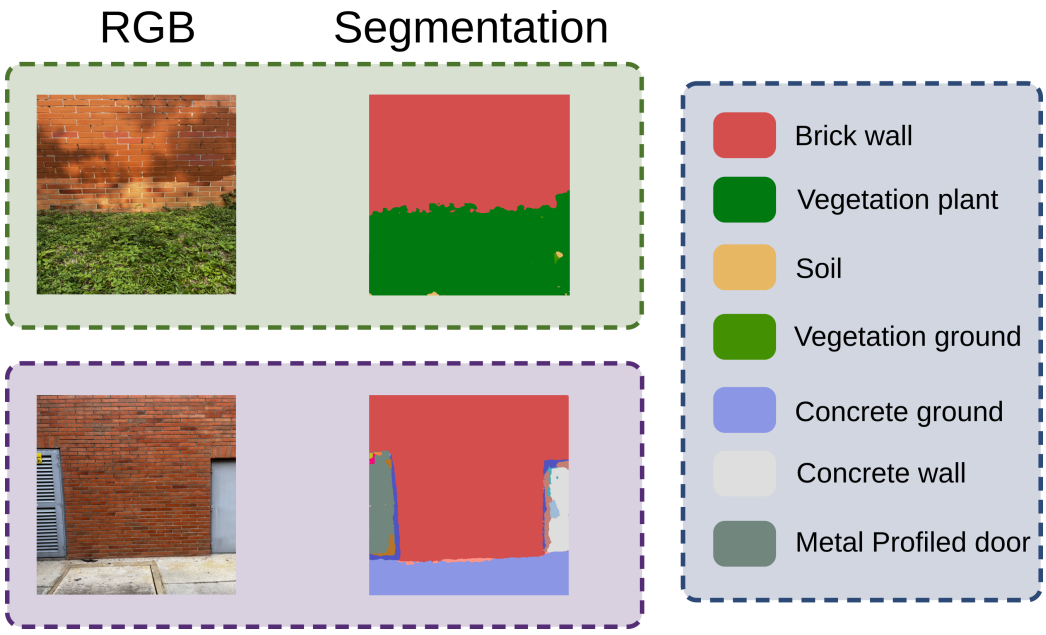

Figura 12. Resultados cualitativos de la predicción segmentación de materiales a partir del método propuesto en imágenes RGB capturadas en el campus universitario. Se observa una correcta identificación y delimitación de los principales materiales presentes en las escenas, demostrando la capacidad de generalización del modelo a entornos no vistos previamente.

6. CONCLUSIONES

En este trabajo se presentó un método novedoso para la segmentación de materiales en imágenes RGB que integra eficientemente información espectral mediante una arquitectura de *transformers* de visión. El enfoque propuesto, centrado en el módulo *Adaptive Spectral Prompts* (ASP), demostró una robustez significativa ante la ausencia de modalidad espectral durante la inferencia, manteniendo un buen rendimiento solo con información RGB. Esta capacidad de adaptación a escenarios de modalidad faltante es particularmente valiosa para aplicaciones del mundo real donde los sensores espectrales pueden no estar siempre disponibles. El método logró superar consistentemente a los enfoques del estado del arte en el conjunto de datos LIB-HSI, tanto en precisión global como en IoU promedio por clase, validando la hipótesis de que la incorporación de información espectral, incluso de manera embebida, puede mejorar significativamente la tarea de segmentación de materiales. Además, la validación cualitativa en imágenes fuera del conjunto de datos de entrenamiento demostró la capacidad del modelo para generalizar efectivamente a nuevos escenarios, sugiriendo su potencial para aplicaciones prácticas diversas. La eficiencia computacional del modelo, lograda a través de la arquitectura ASP y las técnicas de *prompt tuning*, permite su implementación en entornos con recursos limitados, ampliando así su aplicabilidad. Estos resultados subrayan el potencial de los enfoques multimodales adaptivos en la mejora de tareas de visión por computadora complejas, como la segmentación de materiales, y abren caminos prometedores para futuras investigaciones en este campo.

7. TRABAJO FUTURO

El éxito del método propuesto abre varias líneas prometedoras para investigación futura. En primer lugar, se podría explorar la adaptación del modelo a otras modalidades de datos complementarias, como información de profundidad o térmica, para mejorar aún más la precisión en la segmentación de materiales. Asimismo, sería valioso investigar la aplicabilidad del enfoque ASP en otras tareas de visión por computadora, como la detección de objetos o la estimación de pose, para evaluar su versatilidad. En cuanto al método en sí, explorar la integración de aproximaciones de bajo rango para el *fine-tuning* del *decoder* o de los parámetros del módulo ASP podría ofrecer mejoras adicionales en eficiencia computacional y rendimiento. Dada la escasez de conjuntos de datos que combinen información RGB y espectral para la segmentación de materiales, un esfuerzo significativo podría dirigirse a la creación de un nuevo conjunto de datos más amplio y diverso, que incluya una variedad de escenarios, condiciones de iluminación y tipos de materiales. Este recurso no solo beneficiaría directamente nuestra línea de investigación, sino que también podría impulsar avances en el campo de la segmentación de materiales en general. Finalmente, sería beneficioso ampliar la validación del método en condiciones más variadas y desafiantes, para robustecer aún más su generalización a escenarios del mundo real. Estas extensiones podrían consolidar y expandir la utilidad del enfoque propuesto en una variedad de aplicaciones prácticas y contextos de investigación.

BIBLIOGRAFÍA

Adarsh, S et al. “Performance comparison of Infrared and Ultrasonic sensors for obstacles of different materials in vehicle/robot navigation applications”. En: *IOP Conference Series: Materials Science and Engineering*. Vol. 149. 1. IOP publishing. 2016, pág. 012141 (vid. págs. 15, 19, 23).

Adelson, Edward H. “On seeing stuff: the perception of materials by humans and machines”. En: *Human Vision and Electronic Imaging VI*. Vol. 4299. SPIE. 2001, págs. 1-12 (vid. págs. 14, 19).

Akbari, Hassan et al. “Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text”. En: *Advances in Neural Information Processing Systems* 34 (2021), págs. 24206-24221 (vid. pág. 30).

Arnab, Anurag et al. “Joint Object-Material Category Segmentation from Audio-Visual Cues”. En: *Proceedings of the British Machine Vision Conference (BMVC)*. 2015 (vid. pág. 19).

Arnab, Anurag et al. “Vivit: A video vision transformer”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 6836-6846 (vid. pág. 28).

Ba, Jimmy Lei. “Layer normalization”. En: *arXiv preprint arXiv:1607.06450* (2016) (vid. pág. 42).

Bacca, Jorge, Emmanuel Martinez y Henry Arguello. “Computational spectral imaging: A contemporary overview”. En: *JOSA A* 40.4 (2023), págs. C115-C125 (vid. pág. 24).

Bell, Sean et al. “Material recognition in the wild with the materials in context database”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2015, págs. 3479-3487 (vid. págs. 21, 51).

Carion, Nicolas et al. “End-to-end object detection with transformers”. En: *European Conference on Computer Vision*. Springer. 2020, págs. 213-229 (vid. pág. 28).

Chen, Long et al. “Context-Aware Mixed Reality: A Learning-Based Framework for Semantic-Level Interaction”. En: *Computer Graphics Forum*. Vol. 39. 1. Wiley Online Library. 2020, págs. 484-496 (vid. pág. 19).

Chen, Xiaokang et al. “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation”. En: *European Conference on Computer Vision*. Springer. 2020, págs. 561-577 (vid. pág. 30).

Cheng, Bowen, Alex Schwing y Alexander Kirillov. “Per-pixel classification is not all you need for semantic segmentation”. En: *Advances in Neural Information Processing Systems* 34 (2021), págs. 17864-17875 (vid. pág. 16).

Cockburn, Iain M, Rebecca Henderson y Scott Stern. *The impact of artificial intelligence on innovation*. Vol. 24449. National Bureau of Economic Research Cambridge, MA, USA, 2018 (vid. pág. 13).

Csurka, Gabriela et al. “What is a good evaluation measure for semantic segmentation?.” En: *Bmvc*. Vol. 27. 2013. Bristol. 2013, págs. 10-5244 (vid. págs. 51, 53).

Deng, Jia et al. “Imagenet: A large-scale hierarchical image database”. En: *2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Ieee. 2009, págs. 248-255 (vid. pág. 14).

Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. En: *International Conference on Learning Representations*. 2021 (vid. págs. 16, 25, 27, 38).

Du, Zhuoran et al. “Exploring the applicability of spectral recovery in semantic segmentation of RGB images”. En: *IEEE Transactions on Multimedia* (2024) (vid. pág. 59).

Dubey, Abhimanyu et al. “The llama 3 herd of models”. En: *arXiv preprint arXiv:2407.21783* (2024) (vid. pág. 31).

Foster, David H. et al. “Frequency of metamerism in natural scenes”. En: *J. Opt. Soc. Am. A* 23.10 (2006), págs. 2359-2372 (vid. pág. 20).

Gao, Wei et al. “Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection”. En: *IEEE Transactions on Circuits and Systems for Video Technology* 32.4 (2021), págs. 2091-2106 (vid. pág. 30).

Garini, Yuval, Ian T Young y George McNamara. “Spectral imaging: principles and applications”. En: *Cytometry part a: The Journal of the International Society for Analytical Cytology* 69.8 (2006), págs. 735-747 (vid. pág. 23).

Guo, Yanming et al. “A review of semantic segmentation using deep neural networks”. En: *International Journal of Multimedia Information Retrieval* 7 (2018), págs. 87-93 (vid. pág. 14).

Habili, Nariman et al. “A hyperspectral and RGB dataset for building façade segmentation”. En: *European Conference on Computer Vision*. Springer. 2022, págs. 258-267 (vid. págs. 22, 48, 59, 60).

Hassaballah, Mahmoud y Ali Ismail Awad. *Deep learning in computer vision: principles and applications*. CRC Press, 2020 (vid. pág. 13).

Havaei, Mohammad et al. “Hemis: Hetero-modal image segmentation”. En: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, págs. 469-477 (vid. pág. 33).

He, Kaiming et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. En: *Proceedings of the IEEE international conference on computer vision*. 2015, págs. 1026-1034 (vid. pág. 36).

Hendrycks, Dan y Kevin Gimpel. “Gaussian error linear units (gelus)”. En: *arXiv preprint arXiv:1606.08415* (2016) (vid. pág. 44).

Heng, Yuwen et al. “MatSpectNet: Material Segmentation Network with Domain-Aware and Physically-Constrained Hyperspectral Reconstruction”. En: *arXiv preprint arXiv:2307.11466* (2023) (vid. pág. 48).

Hu, Edward J et al. “Lora: Low-rank adaptation of large language models”. En: *arXiv preprint arXiv:2106.09685* (2021) (vid. págs. 41, 43).

Jaegle, Andrew et al. “Perceiver IO: A General Architecture for Structured Inputs & Outputs”. En: *International Conference on Learning Representations*. 2022 (vid. págs. 16, 31).

Jain, Jitesh et al. “Oneformer: One transformer to rule universal image segmentation”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 2989-2998 (vid. pág. 25).

Jia, Menglin et al. “Visual prompt tuning”. En: *European Conference on Computer Vision*. Springer. 2022, págs. 709-727 (vid. págs. 33, 36, 40).

Kaplan, Jared et al. “Scaling laws for neural language models”. En: *arXiv preprint arXiv:2001.08361* (2020) (vid. pág. 27).

Khattak, Muhammad Uzair et al. “Maple: Multi-modal prompt learning”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 19113-19122 (vid. pág. 34).

Kim, Namyup et al. “Restr: Convolution-free referring image segmentation using transformers”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 18145-18154 (vid. pág. 25).

Kirillov, Alexander et al. “Segment anything”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, págs. 4015-4026 (vid. pág. 15).

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. 13).

Kullback, Solomon. *Information theory and statistics*. Courier Corporation, 1997 (vid. pág. 52).

Lee, Yi-Lun et al. “Multimodal prompting with missing modalities for visual recognition”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 14943-14952 (vid. pág. 46).

Lester, Brian, Rami Al-Rfou y Noah Constant. “The power of scale for parameter-efficient prompt tuning”. En: *arXiv preprint arXiv:2104.08691* (2021) (vid. págs. 34, 55).

Li, Yanyu et al. “Efficientformer: Vision transformers at mobilenet speed”. En: *Advances in Neural Information Processing Systems* 35 (2022), págs. 12934-12949 (vid. pág. 28).

Liang, Jingyun et al. “Swinir: Image restoration using swin transformer”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 1833-1844 (vid. pág. 14).

Liang, Yupeng et al. “Multimodal material segmentation”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 19800-19808 (vid. págs. 14, 15, 22).

Lichtman, Jeff W y José-Angel Conchello. “Fluorescence microscopy”. En: *Nature Methods* 2.12 (2005), págs. 910-919 (vid. págs. 15, 23).

Liu, Ce et al. “Exploring features in a bayesian framework for material recognition”. En: *2010 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, págs. 239-246 (vid. pág. 20).

Liu, Pengfei et al. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. En: *ACM Computing Surveys* 55.9 (2023), págs. 1-35 (vid. pág. 33).

Liu, Yuwei, Hongbin Pu y Da-Wen Sun. “Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications”. En: *Trends in Food Science & Technology* 69 (2017), págs. 25-35 (vid. pág. 23).

Liu, Ze et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 10012-10022 (vid. pág. 25).

Loshchilov, I. “Decoupled weight decay regularization”. En: *arXiv preprint arXiv:1711.05101* (2017) (vid. pág. 54).

Lu, Bing et al. “Recent advances of hyperspectral imaging technology and applications in agriculture”. En: *Remote Sensing* 12.16 (2020), pág. 2659 (vid. pág. 23).

Lu, Pan et al. “Learn to explain: Multimodal reasoning via thought chains for science question answering”. En: *Advances in Neural Information Processing Systems* 35 (2022), págs. 2507-2521 (vid. pág. 29).

Ma, Mengmeng et al. “Smil: Multimodal learning with severely missing modality”. En: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, págs. 2302-2310 (vid. pág. 32).

Micikevicius, Paulius et al. “Mixed Precision Training”. En: *International Conference on Learning Representations*. 2018 (vid. pág. 54).

Min, Xiongkuo et al. “A multimodal saliency model for videos with high audio-visual correspondence”. En: *IEEE Transactions on Image Processing* 29 (2020), págs. 3805-3819 (vid. pág. 30).

Ngiam, Jiquan et al. “Multimodal deep learning”. En: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, págs. 689-696 (vid. pág. 29).

O’Mahony, Niall et al. “Deep learning vs. traditional computer vision”. En: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. 2020, págs. 128-144 (vid. pág. 13).

Perez, Fabian y Hoover Rueda-Chacón. “Beyond Appearances: Material Segmentation with Embedded Spectral Information from RGB-D imagery”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2024, págs. 293-301 (vid. pág. 59).

Poulinakis, Konstantinos. *Multimodal Deep Learning: Definition, Examples, Applications*. en. 29 de jul. de 2024. (Visitado 14-09-2024).

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. pág. 35).

Ros, German et al. “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2016, págs. 3234-3243 (vid. pág. 14).

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. 38).

Schwartz, Gabriel y Ko Nishino. “Recognizing material properties from images”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.8 (2019), págs. 1981-1995 (vid. págs. 20, 21).

Shirmard, Hojat et al. “A review of machine learning in processing remote sensing data for mineral exploration”. En: *Remote Sensing of Environment* 268 (2022), pág. 112750 (vid. pág. 23).

Shrestha, Yash Raj, Shiko M Ben-Menahem y Georg Von Krogh. “Organizational decision-making structures in the age of artificial intelligence”. En: *California Management Review* 61.4 (2019), págs. 66-83 (vid. pág. 13).

Srivastava, Siddharth y Gaurav Sharma. “OmniVec2-A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 27412-27424 (vid. pág. 32).

Strudel, Robin et al. “Segmenter: Transformer for semantic segmentation”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 7262-7272 (vid. pág. 28).

Touvron, Hugo et al. “Going deeper with image transformers”. En: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, págs. 32-42 (vid. pág. 27).

Touvron, Hugo et al. “Training data-efficient image transformers & distillation through attention”. En: *International Conference on Machine Learning*. PMLR. 2021, págs. 10347-10357 (vid. pág. 27).

Tsai, Yao-Hung Hubert et al. “Learning Factorized Multimodal Representations”. En: *International Conference on Learning Representations*. 2019 (vid. pág. 33).

Upchurch, Paul y Ransen Niu. “A dense material segmentation dataset for indoor and outdoor scene parsing”. En: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2022, págs. 450-466 (vid. págs. 15, 19, 21).

Vaswani, Ashish et al. “Attention is all you need”. En: *Advances in Neural Information Processing Systems* 30 (2017) (vid. págs. 25, 37, 38, 41, 43).

Wang, Hu et al. “Multi-modal learning with missing modality via shared-specific feature modelling”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, págs. 15878-15887 (vid. pág. 32).

Wang, Wenhai et al. “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 568-578 (vid. pág. 43).

Wei, Zhixiang et al. “Stronger Fewer & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 28619-28630 (vid. pág. 27).

Wu, Penghao y Saining Xie. “V?: Guided Visual Search as a Core Mechanism in Multimodal LLMs”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 13084-13094 (vid. pág. 30).

Wu, Renjie, Hu Wang y Hsiang-Ting Chen. “A comprehensive survey on deep Multimodal Learning with Missing Modality”. En: *arXiv [cs.CV]* (12 de sep. de 2024) (vid. pág. 32).

Xiao, Tete et al. “Unified perceptual parsing for scene understanding”. En: *Proceedings of the European Conference On Computer Vision (ECCV)*. Springer. 2018, págs. 418-434 (vid. pág. 15).

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. págs. 15, 54).

Xu, Peng, Xiatian Zhu y David A Clifton. “Multimodal learning with transformers: A survey”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023), págs. 12113-12132 (vid. pág. 31).

Xue, Jia et al. “Differential angular imaging for material recognition”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017, págs. 764-773 (vid. pág. 22).

Zhai, Xiaohua et al. “Scaling vision transformers”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, págs. 12104-12113 (vid. pág. 27).

Zhang, Yao et al. “mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation”. En: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, págs. 107-117 (vid. pág. 32).

Zhang, Yiyuan et al. “Meta-transformer: A unified framework for multimodal learning”. En: *arXiv preprint arXiv:2307.10802* (2023) (vid. pág. 32).

Zhao, Zhong-Qiu et al. “Object detection with deep learning: A review”. En: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (2019), págs. 3212-3232 (vid. pág. 14).

Zhou, Kaiyang et al. En: *International Journal of Computer Vision* 130.9 (2022), págs. 2337-2348 (vid. pág. 34).

Zhu, Xizhou et al. “Deformable detr: Deformable transformers for end-to-end object detection”. En: *arXiv preprint arXiv:2010.04159* (2020) (vid. pág. 25).