Revista Iberoamericana de Automática e Informática Industrial

Revista Iberoamericana de Automática e Informática Industrial 00 (2026) 1–??

www.revista-riai.org

Método híbrido de localización visual mediante aprendizaje profundo y características analíticas

Marcos Alfaro[a,*], Juan José Cabrera[a], Oscar Reinoso[a,b], Arturo Gil[a], Luis Payá[a,b]

[a] *Instituto Universitario de Investigación en Ingeniería de Elche (I3E), Universidad Miguel Hernández (UMH), Av/ de la Universidad s/n, 03202, Elche, España.*
[b] *Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Camí de Vera, Edificio 3Q, 46020, Valencia, España.*

To cite this article: Alfaro, M., Cabrera, J.J., Reinoso, O., Gil, A., Payá, L. 2025. A hybrid visual localization method using deep learning and hand-crafted features. Revista Iberoamericana de Automática e Informática Industrial 00, 1-5. https://doi.org/10.4995/riai.2020.7133

Resumen

La localización precisa es un desafío fundamental para los robots móviles. Si bien el Reconocimiento Visual de Lugares (VPR) basado en modelos de *Deep Learning* (DL) ha logrado un rendimiento notable, a menudo presenta problemas de generalización y sensibilidad a cambios de iluminación. Este artículo presenta un método híbrido de VPR mediante imágenes omnidireccionales que combina la capacidad de abstracción del DL con la robustez de las características analíticas de bajo nivel (intensidad, gradiente y tono). Se evalúan estrategias de fusión temprana (*early*), intermedia (*middle*) y tardía (*late*) para integrar estos canales auxiliares con la información RGB, utilizando modelos preentrenados sin necesidad de un ajuste fino específico. Los experimentos demuestran que la arquitectura de fusión tardía ofrece el mejor rendimiento. Asimismo, se concluye que la incorporación de características analíticas, concretamente la intensidad y el gradiente, incrementa notablemente la robustez del sistema frente a variaciones de iluminación y entornos no vistos durante el entrenamiento, ofreciendo una solución efectiva y económica para la robótica móvil.

Palabras clave: Robótica móvil, Reconocimiento de lugares, Visión omnidireccional, Fusión sensorial, Aprendizaje profundo.

A hybrid visual localization method using deep learning and hand-crafted features.

Abstract

Precise localization is a fundamental challenge in the navigation of mobile robots. While Visual Place Recognition (VPR) based on deep learning models has achieved remarkable performance, it often faces issues regarding generalization and proneness to substantial appearance shifts. This paper presents a hybrid VPR method using omnidirectional images that combines the abstraction capacity of deep learning with the robustness of low-level handcrafted features (intensity, gradient, and hue). Early, middle and late fusion strategies are evaluated to integrate these auxiliary channels with RGB information, utilizing pre-trained models without specific retraining. Experiments demonstrate that the late fusion techniques yield the best performance. Furthermore, results indicate that incorporating analytical features —specifically intensity and gradient— significantly increases system robustness against lighting variations and unseen environments, providing an effective and low-cost solution for mobile robotics.

Keywords: Mobile robotics, Place recognition, Omnidirectional imaging, Sensory fusion, Deep learning.

1. Introducción

La percepción visual desempeña un papel crucial en la comprensión de escenarios complejos, proporcionando información rica en color, estructura y textura. En el ámbito de la navegación autónoma, el procesamiento de datos capturados por sensores visuales —específicamente cámaras— es fundamental para interpretar el entorno y garantizar una navegación segura. El reciente auge del aprendizaje profundo ha revolucionado este campo, dotando a los sistemas de visión por computador de un rendimiento sin precedentes en tareas críticas como la recons-

*Autor para correspondencia: malfaro@umh.es
Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

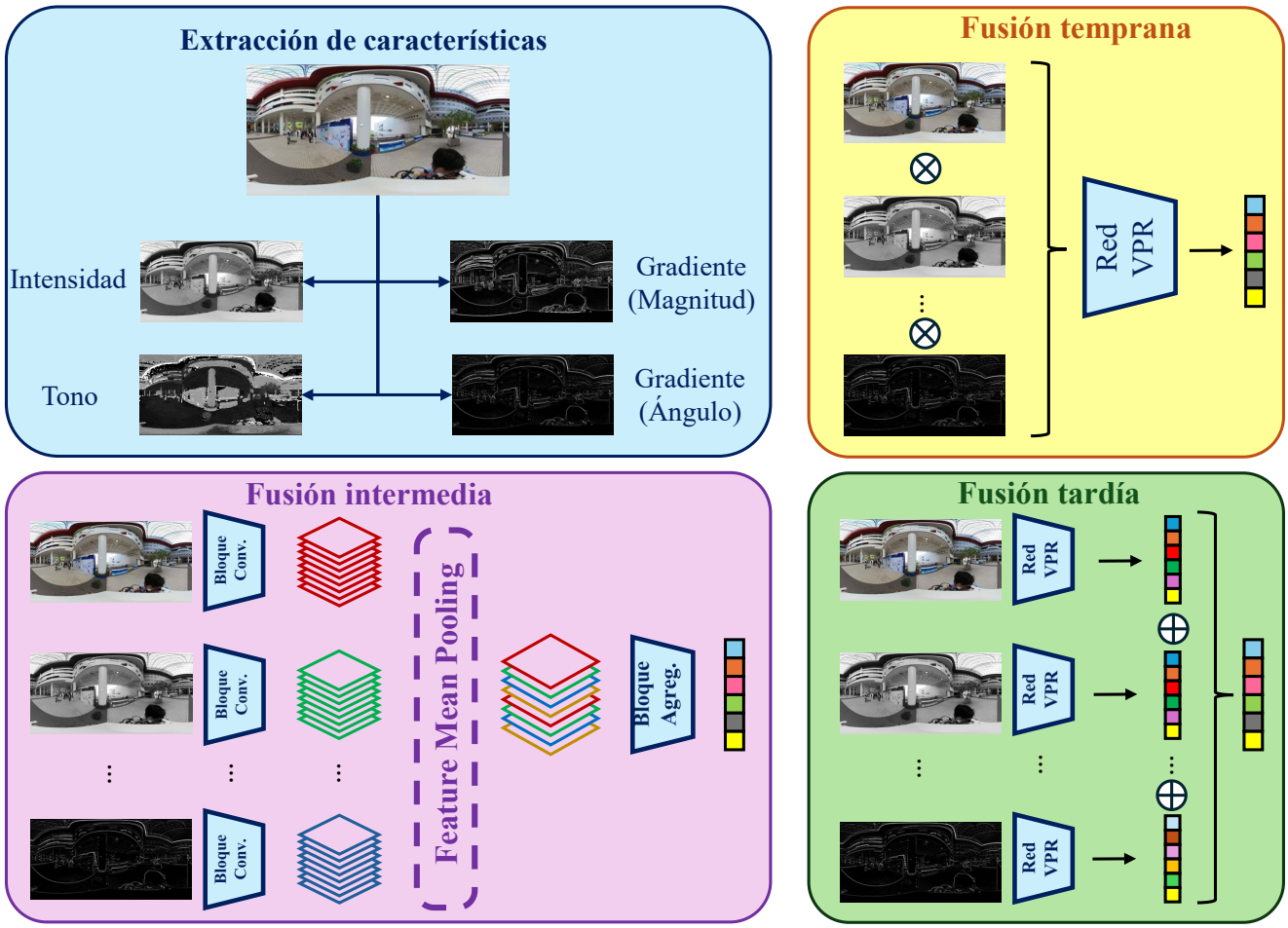


Figura 1: Esquema general del método de VPR propuesto. A partir de imágenes panorámicas, se extraen múltiples representaciones visuales de manera analítica. Posteriormente, se fusionan las imágenes y las características mediante tres técnicas distintas: temprana, intermedia y tardía.

trucción 3D (Chen *et al.*, 2025), la estimación de profundidad (Lin *et al.*, 2025) y la localización (Máximo *et al.*, 2026).

Dentro de la navegación autónoma, la localización se erige como un desafío prioritario, que consiste en determinar la pose (posición y orientación) del vehículo respecto a un mapa del entorno previamente adquirido (Chen *et al.*, 2023). Generalmente, este problema se aborda mediante un enfoque jerárquico: primero, se realiza una estimación aproximada de la ubicación buscando la imagen más similar en el mapa de referencia, tarea conocida como Reconocimiento Visual de Lugares o *Visual Place Recognition* (VPR) (Cabrera *et al.*, 2025). Posteriormente, se ejecuta un refinamiento métrico de la pose mediante correspondencias geométricas o la integración de sensores complementarios (LiDAR, odometría, profundidad). Este trabajo pone el foco en la primera etapa: la tarea de VPR.

Para llevar a cabo el reconocimiento de lugares, es imprescindible codificar el contenido de las imágenes en representaciones compactas. Dependiendo de su alcance, estas características se categorizan en dos grupos: locales, las cuales se extraen a partir de puntos de interés como las esquinas o los bordes (Sarlin *et al.*, 2020); o globales, mediante la obtención de una descripción holística de la escena (Liu *et al.*, 2022). Dada la eficiencia computacional y la capacidad de generalización requerida en entornos extensos, este trabajo adopta un enfoque basado en descriptores globales.

Históricamente, los algoritmos de VPR se han dividido en dos grandes vertientes según la naturaleza de sus descriptores. Por un lado, los métodos clásicos o analíticos emplean reglas predefinidas para extraer las características, basándose en propiedades geométricas o fotométricas (Payá *et al.*, 2016). Por otro lado, los métodos basados en DL aprenden a extraer características de alto nivel de abstracción directamente de los datos (Arandjelovic *et al.*, 2016). Si bien estos últimos han superado el rendimiento de los métodos clásicos, a menudo conllevan un alto coste computacional y son susceptibles al sobreajuste (*overfitting*) o a la pérdida de rendimiento ante cambios de dominio no vistos durante el entrenamiento.

Una estrategia prometedora para mitigar estas limitaciones es la fusión de información. Las arquitecturas de fusión se clasifican habitualmente según la etapa en la que se combinan los datos: fusión temprana, que conlleva la combinación de datos antes del procesamiento por el modelo (Alfaro *et al.*, 2025); fusión intermedia, donde la integración se realiza en las capas intermedias (Guarrasi *et al.*, 2025); y fusión tardía, donde se realiza un procesamiento independiente para cada modalidad y sus salidas se combinan para generar un descriptor final (Lai *et al.*, 2022). Aunque tradicionalmente esto implica el uso de múltiples sensores (fusión multimodal), existe una alternativa eficiente que maximiza el uso de un único sensor: la fusión de diferentes representaciones de la misma entrada visual.

En este contexto, el presente trabajo propone un método híbrido de VPR que combina la capacidad de extraer características abstractas de los modelos basados en DL con la robustez de las características analíticas de bajo nivel (véase la Figura 1). En concreto, se plantea enriquecer la información RGB de imágenes panorámicas con canales auxiliares extraídos analíticamente —tales como el nivel de gris medio, la magnitud y orientación del gradiente, y el tono—, evaluando distintas estrategias de fusión: temprana, intermedia y tardía. De esta forma, el enfoque propuesto permite abordar la localización de manera robusta utilizando cámaras omnidireccionales como única fuente sensorial, ofreciendo una solución efectiva y de bajo coste para la robótica móvil. Las contribuciones principales del presente artículo son las siguientes:

- El desarrollo de un método híbrido de VPR que integra la precisión de descriptores obtenidos por modelos de DL con la información estructural y fotométrica de características analíticas.
- Una comparación exhaustiva entre distintas arquitecturas de fusión temprana, intermedia y tardía aplicadas a imágenes panorámicas y sus características derivadas.
- Una propuesta que demuestra una mejora consistente del rendimiento en diversos entornos y condiciones de iluminación, utilizando modelos preentrenados sin necesidad de un ajuste fino específico.

El resto del artículo se estructura de la siguiente manera: la Sección 2 revisa los trabajos previos en VPR y en métodos de fusión; en la Sección 3 se describe el método propuesto; la Sección 4 presenta las bases de datos utilizadas, los experimentos realizados y los resultados obtenidos; por último, la Sección 5 resume las conclusiones y plantea futuras líneas de investigación.

2. Estado del arte

Esta sección revisa el estado del arte en Reconocimiento Visual de Lugares —Sección 2.1— y en las técnicas de fusión desarrolladas hasta la actualidad —Sección 2.2—.

2.1. Reconocimiento Visual de Lugares

Según el método de extracción de características a partir de las imágenes, los métodos de VPR se pueden separar en dos ramas: los métodos analíticos y los métodos basados en DL.

En cuanto a la extracción de características analíticas, existen métodos que obtienen descriptores de apariencia global a partir del dominio frecuencial como la Firma de Fourier (Menegatti *et al.*, 2004), a partir de la orientación del gradiente como HOG (Zhu *et al.*, 2006) o a partir de filtros de Gabor como, por ejemplo, Gist (Siagian e Itti, 2025). También existen métodos basados en la obtención de descriptores globales a partir de características locales, como *Bag of Words* (BoW) (Gálvez-López y Tardos, 2025).

Debido al incremento de la capacidad computacional de los dispositivos actuales, los métodos basados en DL han experimentado un auge sustancial, desplazando a los métodos tradicionales (Arandjelovic *et al.*, 2016). Algunas arquitecturas clásicas de aprendizaje profundo, como VGG16 (Simonyan y Zisserman, 2014) y ResNet (He *et al.*, 2016), diseñadas originalmente para la clasificación de objetos, han demostrado una excelente capacidad para extraer características. Por ello, modelos de vanguardia en VPR como CosPlace (Berton *et al.*, 2022), utilizado en este trabajo, o EigenPlaces (Berton *et al.*, 2023) aprovechan estos bloques convolucionales como base para generar representaciones compactas y robustas de la escena. Recientemente, han surgido métodos basados en Transformers Visuales, como AnyLoc (Keetha *et al.*, 2023) o SALAD (Izquierdo y Civera, 2024). Mientras que los primeros se caracterizan por una mayor simplicidad y eficiencia computacional, los segundos muestran una mayor capacidad de generalización.

Aunque la visión omnidireccional permite obtener representaciones más completas de la escena respecto a las cámaras estándar para tareas de localización (Wang *et al.*, 2018; Rostkowska y Skrzypczynski, 2023; Cabrera *et al.*, 2026), los métodos mencionados en el párrafo anterior han sido entrenados únicamente con imágenes *pinhole*. Por lo tanto, para extraer el máximo rendimiento a partir de ellos cuando se dispone de imágenes panorámicas es conveniente realizar un ajuste fino de los pesos del modelo.

2.2. Reconocimiento de Lugares Multimodal

Los enfoques multimodales para la tarea de reconocimiento de lugares tratan de mejorar la robustez de los sensores visuales, que son susceptibles a fenómenos como el solapamiento visual o los cambios de apariencia de la escena debidos a cambios lumínicos (Li *et al.*, 2025). Para ello, es común combinarlos con otros tipos de sensores, como por ejemplo el LiDAR (Di Giammarino *et al.*, 2021) o los sensores de odometría (Fan *et al.*, 2025).

En lo que respecta a las técnicas de fusión, algunos trabajos optan por métodos de fusión temprana, ya que proporcionan el contexto completo al modelo (Alfaro *et al.*, 2025). Por el contrario, los métodos de fusión tardía proporcionan una mayor robustez ante la ausencia de la medida de algún sensor (Hausler *et al.*, 2019). Los métodos de fusión intermedia proponen un equilibrio entre ambos enfoques, pero su implementación conlleva una mayor complejidad estructural y opacidad en el modelo respecto a los métodos anteriores (Zhou *et al.*, 2023).
Si bien los métodos de reconocimiento de lugares actuales han demostrado un desempeño y una capacidad de generalización notables, todavía existe margen de mejora en cuanto al balance entre rendimiento y eficiencia. Además, los enfoques multimodales introducen costes y complejidad adicionales. En consecuencia, el presente trabajo busca mejorar este balance mediante una idea de integración directa y de fácil despliegue, pero también robusta ante situaciones desafiantes para el agente móvil, como son los cambios de apariencia debidos a variaciones de iluminación, la presencia de elementos dinámicos o el solapamiento visual, donde los métodos actuales son más propensos a fallar.

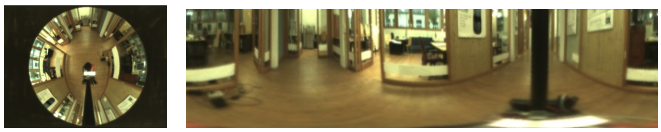
3. Metodología

3.1. Visión omnidireccional

En este trabajo, se ha optado por las cámaras omnidireccionales como única fuente de información sensorial empleada para la tarea de VPR. Esta elección se fundamenta en su elevado

campo de visión que permite obtener representaciones completas de la escena independientemente de la orientación del robot. Concretamente, se ha trabajado con imágenes panorámicas, las cuales pueden obtenerse principalmente por medio de dos sistemas omnidireccionales distintos:

- Sistema de visión catadióptrico: consta de una cámara estándar y un espejo, habitualmente hiperbólico. Se fundamenta en que los rayos de luz, provenientes de todas las direcciones, inciden en el espejo y se proyectan en el foco de la hipérbola, punto donde se sitúa el foco de la cámara. El resultado es una imagen omnidireccional (véase la Figura 2a), a partir de la cual existe una transformación directa al formato panorámico (Figura 2b).
- Cámara 360° (Figura 2c): este tipo de cámaras captura imágenes de alta resolución, con un campo de visión de 360 grados en los ejes vertical y horizontal. Para obtener una imagen equirrectangular, basta con realizar una proyección esférica sobre la imagen (Flores *et al.*, 2024).



(a) Omnidireccional (b) Panorámica



(c) Equirrectangular

Figura 2: (a) Muestra de una imagen omnidireccional de la base de datos COLD (Pronobis y Caputo, 2009), (b) su imagen panorámica correspondiente y (c) muestra de una imagen equirrectangular del dataset 360Loc (Huang *et al.*, 2024).

3.2. Características analíticas visuales

Para abordar la tarea de VPR, este artículo propone un método híbrido que combina modelos de aprendizaje profundo con características que han sido extraídas a partir de las imágenes panorámicas mediante métodos analíticos. En este apartado se describen las características utilizadas. La Figura 3 muestra ejemplos de las distintas características extraídas a partir de una imagen panorámica. A continuación, se da un listado de las características extraídas:

- Intensidad o brillo: consiste en convertir la imagen a escala de grises, promediando el valor de los canales RGB.
- Tono (*Hue*): se trata del color puro que predomina en un píxel determinado, y se expresa mediante un ángulo.
- Gradiente: mide el cambio de intensidad en el entorno de un píxel. A través de máscaras de Sobel de 3x3 píxeles, se obtienen los bordes de la imagen original, y por lo tanto información estructural relevante.

  A partir de la información proporcionada por el gradiente, se pueden obtener dos características distintas:

  - Magnitud: mide cómo de grande es el cambio de intensidad para cada píxel respecto a sus vecinos.
  - Orientación: dirección de la máxima variación de intensidad para cada píxel.

3.3. Modelo de Reconocimiento Visual de Lugares

En este trabajo, se ha hecho uso de CosPlace (Berton *et al.*, 2022) como método de obtención de descriptores para la tarea de reconocimiento de lugares. Los modelos han sido preentrenados con 41,2 millones de imágenes para dicha tarea, lo cual permite obtener resultados muy precisos a pesar de que CosPlace utiliza arquitecturas relativamente sencillas como ResNet (He *et al.*, 2016) o VGG (Simonyan y Zisserman, 2014). Por lo tanto, se trata de una solución precisa y eficiente que puede ser empleada como base para el método de VPR propuesto en este trabajo.

Para la validación experimental, se ha empleado el modelo de CosPlace que utiliza VGG-16 como bloque convolucional y un tamaño de descriptor de 512. Para aprovechar las capacidades adquiridas por este modelo durante su extenso preentrenamiento, en este trabajo se ha tratado de realizar las mínimas modificaciones posibles sobre el modelo. Concretamente, se ha modificado la capa de entrada únicamente para el método de fusión temprana (detallado en la Sección 3.4.1), mientras que para la fusión intermedia se ha intercalado una capa de *pooling* entre el bloque convolucional y el bloque de agregación de características de CosPlace, sin alterar ninguna capa (véase la Sección 3.4.2). Para la fusión tardía, la arquitectura del modelo se ha dejado intacta. Respecto al proceso de ajuste fino, se ha aplicado una estrategia de transferencia de aprendizaje, es decir, se ha partido de los pesos del modelo de CosPlace preentrenado. Durante esta fase de ajuste fino, los pesos del bloque de agregación de características permanecen congelados, limitando el reentrenamiento exclusivamente a las capas del bloque convolucional. La Figura 4 muestra la arquitectura del modelo de CosPlace y las operaciones adicionales que se han realizado en este trabajo.

3.4. Métodos de fusión

En la sección de experimentos, se ha llevado a cabo una evaluación comparativa entre distintas técnicas de fusión, las cuales se han aplicado por separado: fusión temprana, intermedia o tardía.

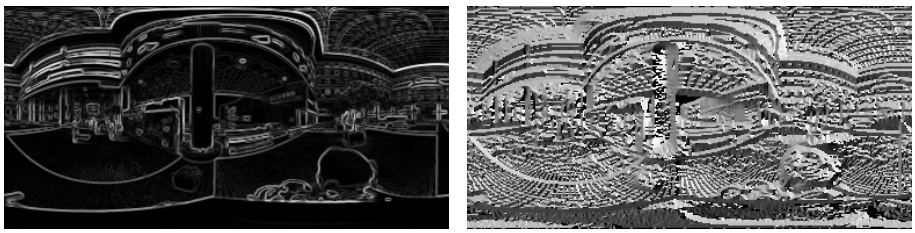
3.4.1. Fusión temprana

La fusión temprana consiste en combinar la información capturada por cada sensor antes de introducirla a la red. De esta manera, un único modelo se encarga de procesar simultáneamente las distintas modalidades.

En este trabajo, la fusión temprana se ha realizado mediante la concatenación de las distintas modalidades en la dimensión del canal. De esta forma, la entrada del modelo tendrá $C = 3+n$ canales, donde n es el número de características añadidas, y por



(a) Imagen Equirrectangular (b) Intensidad (I) (c) Tono



(d) Magnitud del gradiente ($Mag.$) (e) Orientación del gradiente (θ)

Figura 3: (a) Muestra de una imagen equirrectangular de la base de datos 360Loc (Huang et al., 2024) y los mapas de características obtenidos a partir de la imagen: (b) intensidad. (c) tono (hue), (d) magnitud del gradiente y (e) orientación del gradiente.

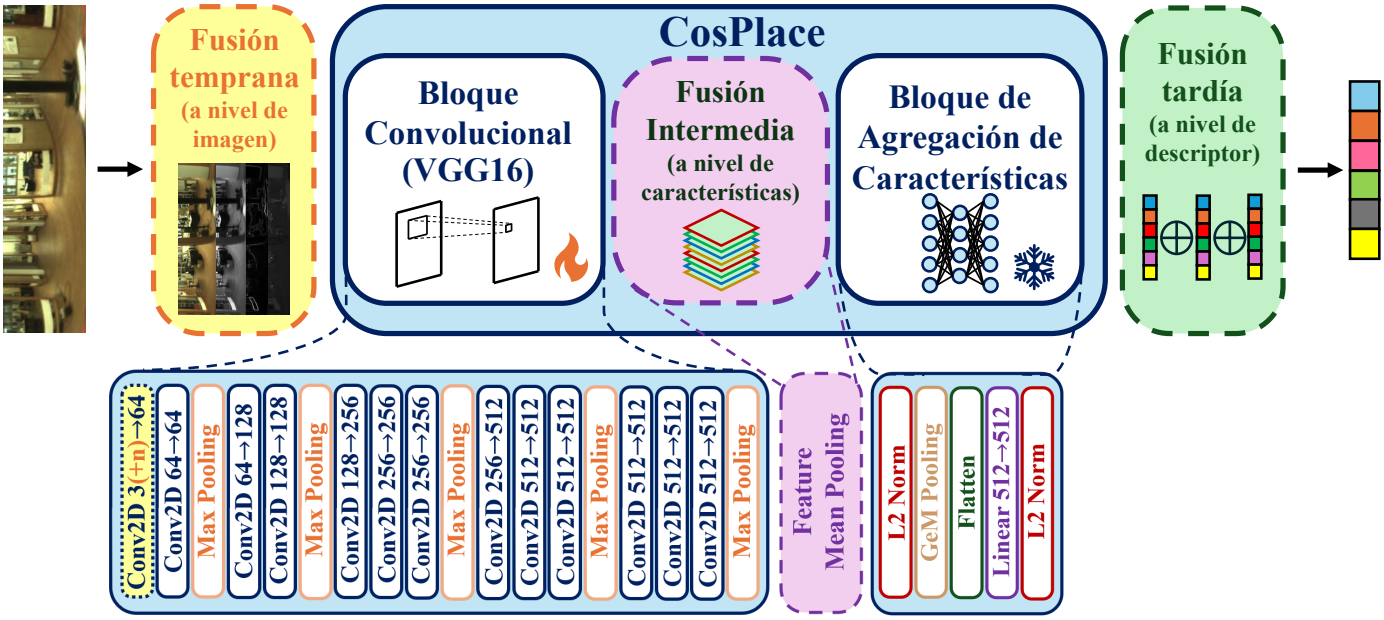


Figura 4: Esquema del método propuesto. Se parte del modelo de CosPlace preentrenado (Berton *et al.*, 2022), formado por un bloque convolucional con la arquitectura de VGG16 (Simonyan y Zisserman, 2014) y un bloque de agregación de características propuesto por (Berton *et al.*, 2022). Sobre este modelo, se aplican tres métodos de fusión que se han evaluado de forma separada: temprana, intermedia y tardía. El bloque de fusión temprana requiere la concatenación de las distintas representaciones visuales a nivel de canal y la adaptación de la primera capa *Conv2D* de $3 \rightarrow 64$ a $3 + n \rightarrow 64$ canales. El bloque de fusión intermedia se intercala entre el bloque convolucional y el bloque de agregación de características de CosPlace. Por último, la fusión tardía se aplica a la salida del modelo, y consiste en la concatenación de los descriptores específicos de cada modalidad. El símbolo del fuego indica los bloques que son reentrenados en este trabajo, el símbolo del copo de nieve indica los bloques que se mantienen congelados, mientras que los bloques de fusión no contienen ningún parámetro aprendible por la red.

lo tanto, la capa de entrada de éste debe ser modificada para aceptar más de 3 canales.

En cuanto a los pesos de la nueva capa de entrada, los canales correspondientes a la imagen RGB no se han modificado, mientras que para los canales adicionales se ha realizado el promedio de los pesos *RGB*.

3.4.2. *Fusión intermedia*

La fusión intermedia trata de agregar la información proveniente de distintas modalidades en capas intermedias de la red. Esto es, se fusionan los mapas de características extraídos por los *encoders* o codificadores adaptados a cada modalidad.

El método propuesto en este trabajo reentrena un bloque convolucional de CosPlace para cada representación visual,

dando lugar a n mapas de características $\mathbf{F}_k$ de dimensiones $C_{out} \times \frac{H}{16} \times \frac{W}{16}$, siendo $C_{out} = 512$ el número de características devuelto por el bloque convolucional de CosPlace y $\mathbf{F}_k$ el mapa de características de la k-ésima modalidad. A continuación, se realiza el promedio de las características de todas las modalidades, obteniendo un mapa de características combinado $\mathbf{F}$ con las mismas dimensiones que $\mathbf{F}_k$. Por último, el mapa de características $\mathbf{F}$ es procesado por el bloque de agregación de características de CosPlace sin ajuste fino, obteniendo un único descriptor global.

3.4.3. *Fusión tardía*

Estos métodos se caracterizan por procesar cada modalidad sensorial por separado, típicamente mediante un modelo independiente para cada modalidad. Posteriormente, se fusionan sus salidas o descriptores para obtener un único descriptor de apariencia global.

En este experimento, un modelo de CosPlace ha sido reentrenado con imágenes de cada modalidad, los cuales son utilizados para realizar la inferencia durante la evaluación. No obstante, también se han incluido resultados sin ajuste fino. Para realizar la fusión tardía, se ha empleado la concatenación de los descriptores de las distintas características.

4. Experimentos

4.1. *Bases de datos utilizadas*

Para evaluar el desempeño del método propuesto en la Sección 3 en la tarea de reconocimiento de lugares, se han empleado dos bases de datos: la base de datos COLD, capturada en distintos entornos de interior, y la base de datos 360Loc, que contiene partes de interior y de exterior.

4.1.1. *COLD*

La base de datos COLD (Pronobis y Caputo, 2009) contiene imágenes omnidireccionales capturadas por robots móviles en distintos edificios de las universidades de Friburgo y Saarbrücken. El sensor empleado es un sistema de visión catadióptrico, compuesto por una cámara estándar y un espejo hiperbólico. Las trayectorias obtenidas pertenecen a cuatro entornos distintos: Friburgo A (*FR-A*) y B (*FR-B*) y Saarbrücken A (*SA-A*) y B (*SA-B*) y han sido capturadas bajo distintas condiciones de iluminación externas (soleado, nublado o noche) que tienen un impacto directo y severo en la apariencia de las escenas debido a la luz natural que penetra a través de grandes ventanales y a la variabilidad de la iluminación artificial. Estos factores introducen desafíos como reflejos especulares, sombras dinámicas y cambios en el balance de blancos que convierten a COLD en una base de datos muy útil para evaluar la robustez del método frente a cambios de iluminación en entornos interiores. Además, las imágenes contienen otros ejemplos desafiantes para la red, tales como oclusiones, ruido o la presencia de elementos dinámicos y solapamiento visual. Por estos motivos, se trata de una base de datos especialmente útil para evaluar el funcionamiento del método propuesto en entornos interiores.

En este trabajo, se ha realizado una proyección cilíndrica sobre las imágenes omnidireccionales, de dimensiones 640 × 480 píxeles, para convertirlas a formato panorámico, obteniendo imágenes de dimensiones 128 × 512 píxeles. Con el fin de evaluar la robustez del método propuesto frente a cambios lumínicos y su capacidad de generalización, los modelos han sido reentrenados con imágenes capturadas en un entorno y condición de iluminación específicos, mientras que han sido evaluados en entornos y condiciones de iluminación diferentes. Para el ajuste fino de la red, se ha empleado el 20 % de las imágenes de la secuencia *seq2_cloudy3* como conjunto de entrenamiento y otro 20 % de imágenes diferentes a las anteriores como conjunto de validación, que es utilizado como criterio para guardar el mejor modelo obtenido durante el entrenamiento. Los conjuntos de entrenamiento y validación no contienen imágenes en común. El conjunto de imágenes consulta (test) está formado por 10 secuencias de los 4 entornos disponibles de COLD (*FR-A*, *FR-B*, *SA-A* y *SA-B*), capturadas bajo las condiciones *nublado*, *noche* y *soleado*, tal y como se muestra en la Tabla 1. Cabe destacar que estas secuencias no tienen ninguna imagen en común con la secuencia de entrenamiento ni con la de validación. Finalmente, durante la evaluación de los modelos se utiliza como modelo visual o mapa un conjunto de imágenes capturado en condiciones de iluminación *nublado*. Este conjunto coincide con la secuencia de entrenamiento en *FR-A*.

Tabla 1: Número de imágenes que componen los conjuntos de entrenamiento, validación y evaluación en la base de datos COLD. #, # y # hacen referencia al número de imágenes capturadas bajo las condiciones *nublado*, *noche* y *soleado*, respectivamente.

Entorno	Entrenamiento	Val.	Evaluación	
			Mod. visual	Consulta
FR-A	556	556	556	2595/2707/2114
FR-B	–	–	560	2008/0/1797
SA-A	–	–	586	2774/2267/0
SA-B	–	–	321	836/870/872

4.1.2. *360Loc*

El dataset 360Loc (Huang *et al.*, 2024) contiene imágenes equirrectangulares que han sido capturadas mediante una cámara 360°. Este dataset está formado por distintas trayectorias pertenecientes a la universidad de Hong Kong, que contienen entornos de interior y exterior, dentro de los cuales se encuentran espacios confinados, parcialmente abiertos y totalmente abiertos. Asimismo, las imágenes han sido capturadas tanto de día como de noche, por lo que se pueden apreciar cambios de apariencia muy pronunciados.

Tabla 2: Número de imágenes que componen los conjuntos de entrenamiento, validación y evaluación en la base de datos 360Loc. # y # hacen referencia al número de imágenes capturadas bajo las condiciones *día* y *noche*, respectivamente.

Entorno	Entrenamiento	Val.	Evaluación	
			Mod. Visual	Consulta
atrium	581	197	581	875/1219
concourse	–	–	491	593/514
hall	–	–	540	1123/1061
piatrium	–	–	632	1008/697

Para llevar a cabo los experimentos, se han redimensionado las imágenes equirrectangulares, originalmente de 6144x3072 píxeles, hasta 384x192 píxeles, con el fin de reducir sustancial-

mente el tiempo de inferencia del modelo. Para el ajuste fino, se ha utilizado la secuencia *database* de la trayectoria *atrium*, mientras que el test se ha realizado en las cuatro trayectorias presentes en el dataset: *atrium*, *concourse*, *hall* y *piatrium*, en condiciones de día y de noche. La Tabla 2 muestra el número de imágenes que componen los conjuntos de entrenamiento y test. Al igual que en COLD, los conjuntos de entrenamiento, validación y test son diferentes y no comparten ninguna imagen. En el caso del entorno *atrium*, el conjunto de entrenamiento se utiliza también como modelo visual.

4.2. Detalles de implementación

Con el fin de asegurar la coherencia con respecto a trabajos previos, se ha adoptado el procedimiento seguido en (Alfaro *et al.*, 2025) para el ajuste fino de los modelos. Los modelos han sido reentrenados con una muestra de 50.000 tripletas extraídas exclusivamente del conjunto de entrenamiento (sin que los datos de validación, test o el mapa visual intervengan en este proceso). Cada tripleta aleatoria está formada por una imagen ancla, una positiva (distancia $\leq r_+$) y una negativa (distancia $\geq r_-$), cumpliéndose siempre que $r_+ \leq r_-$. Dado el tamaño acotado del conjunto de entrenamiento, el muestreo se realiza con reemplazo, permitiendo la reutilización de imágenes en diferentes combinaciones. No obstante, el espacio combinatorio disponible asegura la diversidad del entrenamiento y minimiza la probabilidad de existencia de tripletas idénticas duplicadas, previniendo el sobreajuste. En este trabajo, los umbrales r_+ y r_- han tomado un valor de 0,4m para la base de datos COLD y de $1m$ para la base de datos 360Loc, respectivamente. La función de pérdida utilizada ha sido la Batch Hard Loss (Hermans *et al.*, 2017) y se ha empleado el algoritmo de optimización *Stochastic Gradient Descent* (SGD) con una tasa de aprendizaje lr=0,0001. La única diferencia, tal y como se ha comentado en la Sección 3.3, se basa en que en este trabajo se ha reentrenado únicamente el bloque convolucional de la red, dejando congelados los pesos del bloque de agregación de características.

Para la evaluación, se realiza una inferencia sobre el modelo reentrenado con todas las imágenes del modelo visual y de cada secuencia de consulta (test), obteniendo un descriptor global por imagen. Posteriormente, para cada imagen de test, se compara su descriptor con los descriptores de las imágenes que componen el modelo visual mediante la distancia euclídea. El resultado es una lista con los índices del modelo visual ordenados de menor a mayor distancia euclídea, siendo el primer índice la imagen predicha por el modelo como la más cercana.

El código asociado a este trabajo de investigación se encuentra disponible en el siguiente enlace: `https://github.com/MarcosAlfaro/VPR_HandcraftedFeatures`.

4.3. Métricas utilizadas

Las métricas que se han empleado para la evaluación son el Recall@1 ($R@1$) y el Recall@1 % ($R@1\,\%$). Estas medidas se obtienen a partir del error de localización e_j para la imagen j de test, definido como la distancia geométrica entre las coordenadas de la imagen de test y la imagen del modelo visual que ha sido predicha como la más cercana (i.e. la primera candidata recuperada), esto es:

$$e_j = \left\| (x_{test}, y_{test}) - (x_{pred}, y_{pred}) \right\|_2 \quad (1)$$

A partir de este error, el $R@1$ se define como la proporción de imágenes de test para las cuales la primera candidata recuperada cumple la condición $e_j \leq d$, siendo d un umbral que depende de la base de datos utilizada. Por su parte, la métrica $R@1\,\%$ contabiliza un acierto si, entre las K primeras candidatas recuperadas (siendo K el 1 % del número de imágenes del modelo visual) existe al menos una que satisface $e_j \leq d$. Teniendo en cuenta la distancia media entre puntos de captura consecutivos en cada conjunto de datos, en este trabajo se ha empleado un umbral d=0,5m para todos los entornos de COLD, $d = 5m$ para el entorno *concourse* de 360Loc, y $d = 10m$ para el resto de entornos de 360Loc.

4.4. Análisis experimental

4.4.1. Resultados sin fusión

En primer lugar, se aborda la tarea de VPR con cada una de las características visuales por separado, es decir, no se realiza ningún tipo de fusión. Las Tablas 3 y 4 muestran los resultados globales de $R@1$ y $R@1\,\%$ para todos los entornos y condiciones de iluminación en COLD y 360Loc, respectivamente. Los mejores resultados se muestran en **negrita**, mientras que los segundos mejores resultados aparecen subrayados.

Tabla 3: Rendimiento del modelo de CosPlace en COLD con cada una de las características visuales por separado, sin ajuste fino (izquierda) y con ajuste fino (derecha).

COLD	sin ajuste fino		con ajuste fino	
Característica	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	**84,00**	94,85	85,70	96,20
Intensidad (I)	82,87	**94,87**	**85,82**	**96,21**
Tono	59,55	79,63	70,70	85,96
Gradiente (Mag)	77,84	92,70	82,82	95,23
Gradiente (θ)	66,32	86,57	78,79	92,55

Tabla 4: Rendimiento del modelo de CosPlace en 360Loc con cada una de las características visuales por separado, sin ajuste fino (izquierda) y con ajuste fino (derecha).

360Loc	sin ajuste fino		con ajuste fino	
Característica	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	**80,62**	**89,46**	**82,55**	**90,62**
Intensidad (I)	76,46	85,97	79,79	88,37
Tono	49,57	71,87	55,64	67,61
Gradiente (Mag)	76,62	88,66	79,28	90,08
Gradiente (θ)	61,51	79,05	69,33	83,84

Las Tablas 3 y 4 muestran que, como era previsible, se obtienen mejores resultados con las imágenes estándar (RGB), ya que el modelo ha sido preentrenado con millones de imágenes de este tipo. No obstante, cuando se realiza un ajuste fino, algunas características como la magnitud del gradiente o la intensidad permiten alcanzar resultados similares a los obtenidos con imágenes a color. Estos resultados dan pie a explorar la fusión de distintas características visuales para realizar el reconocimiento de lugares. Las Secciones 4.4.2, 4.4.3 y 4.4.4 recogen los resultados obtenidos mediante técnicas de fusión temprana, intermedia y tardía, respectivamente.

4.4.2. Fusión temprana

A continuación, se ha realizado una fusión temprana entre las características visuales. Esta fusión se ha realizado mediante una concatenación a nivel de canal, siguiendo el procedimiento detallado en la Sección 3.4.1. Las Tablas 5 y 6 muestran el $R@1$ y el $R@1\,\%$ global con diferentes conjuntos de características en COLD y en 360Loc, respectivamente.

Tabla 5: Resultados de $R@1$ y $R@1\,\%$ obtenidos mediante la fusión temprana de distintos conjuntos de características en la base de datos COLD, sin ajuste fino (izquierda) y con ajuste fino (derecha).

COLD	sin ajuste fino		con ajuste fino	
Características	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	84,00	94,85	85,70	96,20
I	82,87	94,87	**85,82**	96,21
RGB + I	**84,32**	**95,41**	85,78	**96,37**
RGB + Tono	77,74	92,33	80,85	94,00
RGB + Mag	80,20	93,71	83,06	95,18
RGB + θ	75,33	90,61	81,53	94,30
RGB + I + Mag	80,96	94,39	83,63	95,63
Todas exc. Tono	79,71	93,25	84,18	95,92
Todas	79,81	93,08	84,51	95,70

Tabla 6: Resultados de $R@1$ y $R@1\,\%$ obtenidos mediante la fusión temprana de distintos conjuntos de características en la base de datos 360Loc, sin ajuste fino (izquierda) y con ajuste fino (derecha).

360Loc	sin ajuste fino		con ajuste fino	
Características	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	**80,62**	**89,46**	82,55	90,62
I	76,46	85,97	79,79	88,37
RGB + I	80,53	89,32	82,68	90,77
RGB + Tono	77,80	87,99	79,57	87,35
RGB + Mag	78,94	88,30	82,36	90,41
RGB + θ	77,02	87,02	79,18	86,93
RGB + I + Mag	79,54	88,87	**82,90**	**91,00**
Todas exc. Tono	80,53	89,24	79,83	87,64
Todas	78,55	89,08	80,79	88,69

En cuanto a la fusión temprana, en entornos de COLD (véase la Tabla 5), combinar las imágenes RGB y la intensidad proporciona los mejores resultados cuando no se realiza el ajuste fino. Después del ajuste fino, los mejores resultados se obtienen con el modelo reentrenado solo con la intensidad (I). En los entornos de 360Loc (véase la Tabla 6), las imágenes RGB originales devuelven el mayor $R@1$ sin ajuste fino, mientras que, tras el ajuste fino, la configuración óptima consiste en la fusión de imágenes RGB, intensidad y magnitud del gradiente. No obstante, ningún proceso ha devuelto una mejora sustancial si lo comparamos con el caso base en el que no se realiza fusión.

4.4.3. Fusión intermedia

Esta sección contiene los resultados obtenidos para la fusión intermedia en COLD y en 360Loc (véase Tablas 7 y 8). En este trabajo, la fusión intermedia se ha llevado a cabo mediante el promedio de los mapas de características de cada representación visual, obtenidos a partir de la salida devuelta por el bloque convolucional de cada modalidad. El procedimiento se encuentra detallado en la Sección 3.4.2.

Tabla 7: Resultados de $R@1$ y $R@1\,\%$ obtenidos mediante la fusión intermedia de distintos conjuntos de características en la base de datos COLD, sin ajuste fino (izquierda) y con ajuste fino (derecha).

COLD	sin ajuste fino		con ajuste fino	
Características	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	84,00	94,85	85,70	96,20
I	82,87	94,87	85,82	96,21
RGB + I	84,15	95,55	**86,25**	96,30
RGB + Tono	83,39	95,04	82,34	95,22
RGB + Mag	83,83	95,46	85,53	96,02
RGB + θ	82,67	95,71	84,71	96,14
RGB + I + Mag	**84,19**	95,60	86,01	96,25
Todas exc. Tono	83,88	**95,84**	85,94	96,41
Todas	83,95	95,80	85,82	**96,57**

Tabla 8: Resultados de $R@1$ y $R@1\,\%$ obtenidos mediante la fusión intermedia de distintos conjuntos de características en la base de datos 360Loc, sin ajuste fino (izquierda) y con ajuste fino (derecha).

360Loc	sin ajuste fino		con ajuste fino	
Características	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	80,62	89,46	82,55	90,62
I	76,46	85,97	79,79	88,37
RGB + *I*	80,35	88,70	83,70	90,62
RGB + Tono	80,61	89,46	83,63	91,10
RGB + *Mag*	81,98	**90,28**	84,70	91,79
RGB + *θ*	79,99	89,16	83,25	91,60
RGB + *I* + *Mag*	**83,13**	89,93	**85,12**	**91,88**
Todas exc. Tono	81,97	90,25	84,82	91,70
Todas	81,86	89,47	84,59	91,61

Los resultados de la Tabla 7 muestran que las combinaciones que incluyen la intensidad ($RGB + I$ y $RGB + I + Mag$) han obtenido los mejores resultados en la base de datos COLD. En cambio, las combinaciones que incluyen la magnitud del gradiente han tenido un mejor desempeño en 360Loc, de acuerdo con la Tabla 8. En estos casos, se puede apreciar una mejora respecto al modelo que ha sido entrenado únicamente con RGB. Esta mejora es especialmente significativa en 360Loc, situada en torno al 2,5 % incluso antes del ajuste fino.

4.4.4. Fusión tardía

En este apartado, se ha llevado a cabo la fusión tardía con los mismos conjuntos de características evaluadas en el apartado anterior. El método de fusión empleado ha sido la concatenación a nivel de descriptor, y el procedimiento seguido se encuentra detallado en la Sección 3.4.3. Las Tablas 9 y 10 muestran los resultados en términos de $R@1$ y $R@1\,\%$ con cada uno de los conjuntos de características en las bases de datos COLD y 360Loc, respectivamente.

A partir de las Tablas 9 y 10 se puede observar que realizar una fusión tardía entre imágenes panorámicas y las características estudiadas permite mejorar de forma notable los resultados de $R@1$ y $R@1\,\%$ en ambas bases de datos, incluso sin realizar ningún ajuste fino específico. Además, cuando se realiza un ajuste fino sobre el modelo de CosPlace con estas características, la diferencia es todavía mayor. En términos generales, han producido resultados destacables tanto la fusión de imágenes RGB con intensidad y magnitud del gradiente como el uso de todas las características excepto el tono.

Tabla 9: Resultados de $R@1$ y $R@1\,\%$ obtenidos mediante la fusión tardía de distintos conjuntos de características en la base de datos COLD, sin ajuste fino (izquierda) y con ajuste fino (derecha).

COLD	sin ajuste fino		con ajuste fino	
Características	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	84,00	94,85	85,70	96,20
I	82,87	94,87	85,82	96,21
RGB + *I*	84,50	95,41	87,08	96,59
RGB + Tono	80,81	84,74	70,70	96,04
RGB + *Mag*	84,12	95,46	86,41	96,40
RGB + θ	82,79	94,95	86,41	96,28
RGB + *I* + *Mag*	**84,69**	**95,62**	87,06	96,59
Todas exc. Tono	84,45	**95,62**	**87,42**	96,76
Todas	83,93	95,33	87,22	**96,80**

Tabla 10: Resultados de $R@1$ y $R@1\,\%$ obtenidos mediante la fusión tardía de distintos conjuntos de características en la base de datos 360Loc, sin ajuste fino (izquierda) y con ajuste fino (derecha).

360Loc	sin ajuste fino		con ajuste fino	
Características	R@1 (%)	R@1 % (%)	R@1 (%)	R@1 % (%)
RGB	80,62	89,46	82,55	90,62
I	76,46	85,97	79,79	88,37
RGB + *I*	80,02	88,13	84,20	90,71
RGB + Tono	75,58	87,49	78,57	88,77
RGB + *Mag*	83,63	91,01	84,72	91,78
RGB + θ	81,64	**91,09**	83,21	91,85
RGB + *I* + *Mag*	82,61	90,04	**85,54**	91,88
Todas exc. Tono	**83,69**	90,99	85,42	**92,07**
Todas	82,90	90,09	85,36	91,79

4.4.5. *Evaluación ante cambios de iluminación*

Con el fin de profundizar en el estudio de la capacidad del método propuesto, se ha analizado su robustez frente a cambios de iluminación del entorno. Para ello, las Figuras 5 y 6 muestran los resultados de $R@1$ en COLD y 360Loc, respectivamente, de las mejores configuraciones de fusión temprana (*RGB* + *I*), intermedia (*RGB*+*I*+*Mag*), tardía (*RGB*+*I*+*Mag*) y del caso sin fusión. En estas figuras, las líneas discontinuas indican el valor medio de R@1 considerando todas las condiciones de iluminación en la evaluación.

A partir de las Figuras 5 y 6 se pueden extraer conclusiones similares. Por un lado, realizar una fusión temprana y no aplicar ningún tipo de fusión presenta un rendimiento global similar. Por otro lado, con la fusión tardía se han obtenido mejores resultados de manera consistente en prácticamente todas las condiciones de iluminación. En cuanto a la fusión intermedia, ha demostrado un mejor desempeño con respecto a la fusión temprana, pero peor que la tardía. Esta diferencia ha sido menor en 360Loc, especialmente bajo condiciones nocturnas, donde ambos métodos han tenido prácticamente el mismo $R@1$ (78,96 % para la fusión tardía, 78,93 % para la fusión intermedia). Asimismo, se puede apreciar que todos los métodos han tenido un rendimiento inferior bajo condiciones nocturnas, lo cual es lógico, ya que esta condición no se ha empleado para el ajuste fino de los modelos. No obstante, la fusión tardía ha mostrado una robustez superior comparada con el resto de métodos, permitiendo reducir la brecha respecto al desempeño bajo condiciones diurnas.

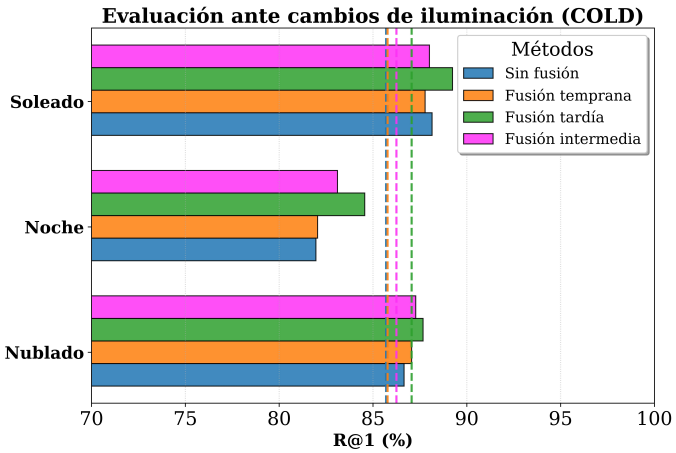


Figura 5: Rendimiento de los distintos métodos (sin fusión, fusión temprana, intermedia y tardía) bajo distintas condiciones de iluminación en la base de datos COLD.

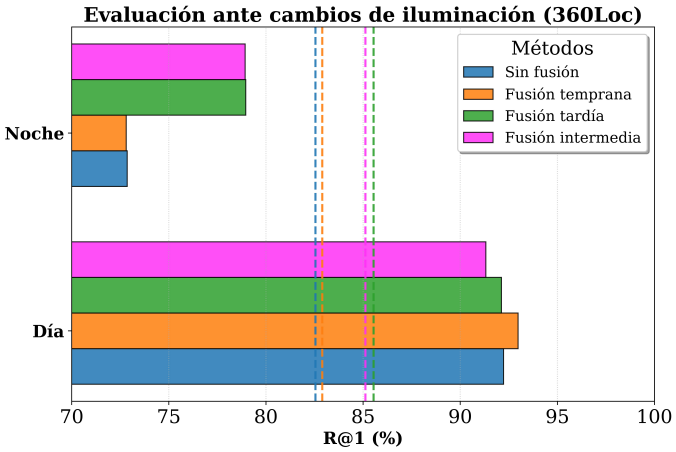


Figura 6: Rendimiento de los distintos métodos (sin fusión, fusión temprana, intermedia y tardía) bajo distintas condiciones de iluminación en la base de datos 360Loc.

4.4.6. *Comparación con el estado del arte*

Para poner en contexto el desempeño del método propuesto en la tarea de VPR, se ha realizado una comparación entre distintos métodos del estado del arte. La Tabla 11 incluye el $R@1$ global obtenido por CosPlace (Berton *et al.*, 2022), EigenPlaces (Berton *et al.*, 2023), AnyLoc (Keetha *et al.*, 2023) y SALAD (Izquierdo y Civera, 2024) en los datasets COLD y 360Loc, así como sus tiempos de inferencia y los tamaños de sus descriptores.

Tabla 11: Comparación entre distintos métodos del estado del arte en los datasets COLD y 360Loc.

Método	R@1 COLD	R@1 360Loc	t. inferencia (ms)	tam. descriptor
CosPlace	84,00	80,62	**0,4**	**512**
EigenPlaces	84,22	78,27	**0,4**	**512**
AnyLoc	77,08	77,99	34,9	1536
SALAD	82,85	**84,88**	9,5	8448
Método propuesto (fusión tardía)	**84,69**	83,69	1,2	1536

La Tabla 11 muestra que el método propuesto ha obtenido el mejor resultado en la base de datos COLD y el segundo mejor resultado en 360Loc. En comparación con SALAD, el método

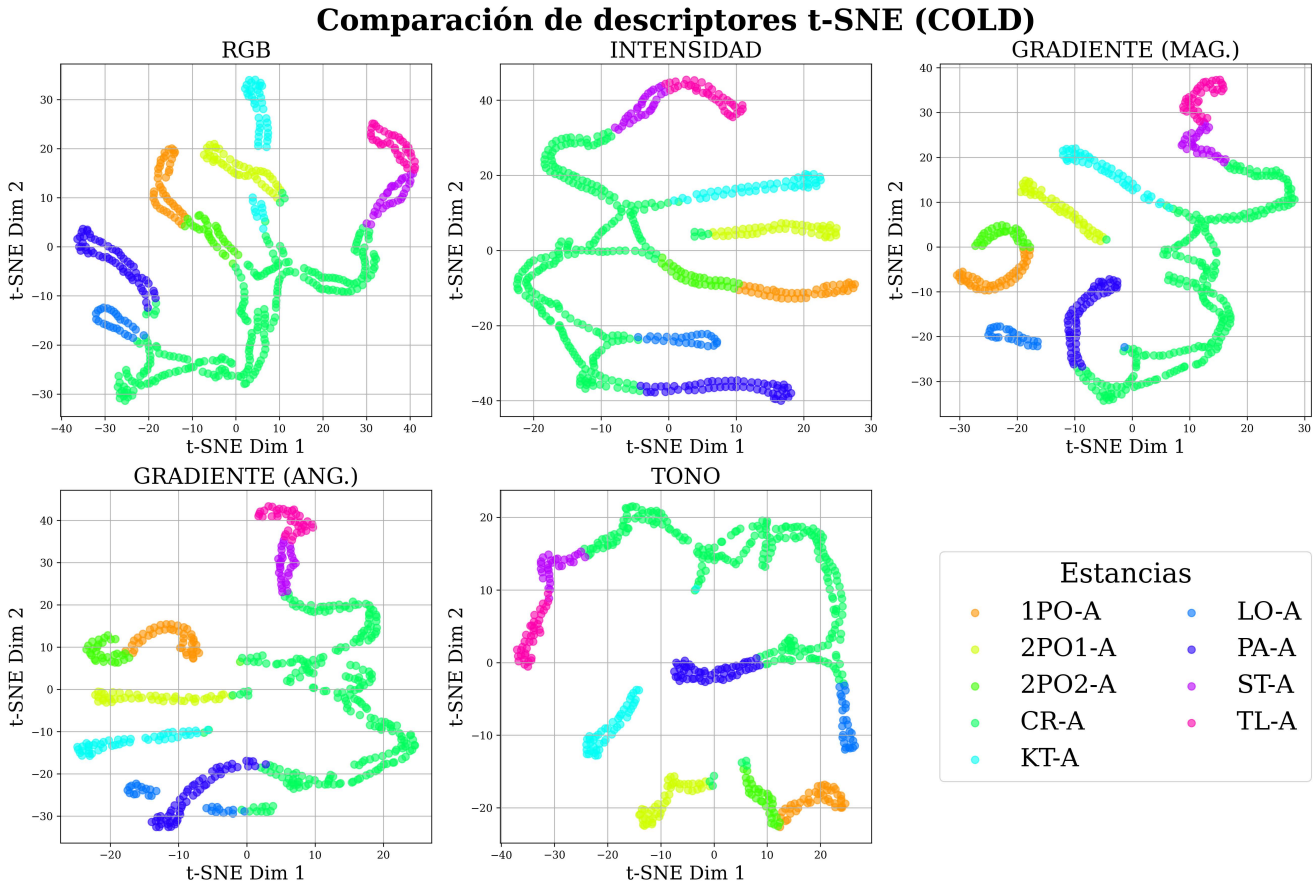


Figura 7: Distribución de los descriptores del modelo visual de *FR-A* (COLD) tras aplicar t-SNE ($n = 2$) con cada una de las características evaluadas. Cada estancia del entorno se ha representado mediante un color distinto.

con los resultados más similares, nuestro método es alrededor de 8 veces más rápido y utiliza un descriptor con un tamaño 5,5 veces menor. En definitiva, estos resultados refuerzan la elección de CosPlace como base de la arquitectura propuesta ya que su combinación con la fusión tardía ofrece un buen compromiso entre rendimiento, tiempo de inferencia y tamaño del descriptor. El resultado de esta combinación es el método que presenta el mejor balance entre rendimiento y eficiencia computacional.

4.5. Evaluación cualitativa

A continuación, se ha realizado una evaluación cualitativa en búsqueda de una mayor comprensión del método propuesto. En la Sección 4.5.1 se ha llevado a cabo un análisis t-SNE, mientras que en la Sección 4.5.2 se ha realizado una discusión sobre las limitaciones del método propuesto.

4.5.1. Calidad de los descriptores obtenidos

En este apartado, se ha realizado un análisis de la calidad de los descriptores obtenidos a partir de cada una de las representaciones visuales. Para ello, se ha empleado la herramienta t-SNE (Maaten y Hinton, 2008), la cual permite visualizar la distribución de un conjunto de descriptores multidimensionales (en nuestro caso $D = 512$) a través de la técnica de Análisis de Componentes Principales (PCA), representando los dos primeros componentes en una gráfica bidimensional.

En este trabajo, se ha empleado t-SNE para visualizar la distribución de un conjunto de descriptores obtenidos a partir de cada una de las características visuales —RGB, intensidad, tono, magnitud y orientación del gradiente— y su modelo reentrenado correspondiente. La Figura 7 muestra el resultado de aplicar t-SNE a los descriptores del modelo visual del entorno Friburgo A (*FR-A*) de la base de datos COLD. Se ha optado por esta base de datos ya que se encuentra dividida por estancias y por lo tanto resulta más sencilla la separación de las imágenes en distintas clases. Se han utilizado colores diferenciados para representar cada una de las estancias del entorno.

La Figura 7 revela una distribución consistente con la estructura del entorno para todas las características. Los descriptores correspondientes al pasillo (*CR-A*) ocupan una posición central, con varias ramificaciones que representan las distintas habitaciones. También se puede apreciar que los modelos tienden a agrupar las estancias semánticamente similares, como es el caso de las oficinas (*1PO-A*, *2PO1-A*, *2PO2-A* y *LO-A*). Aún así, los descriptores correspondientes a una misma estancia suelen formar un grupo claramente diferenciado, lo que demuestra que el modelo es capaz de distinguir entre habitaciones similares. En cuanto a la comparación de las distintas características, se pueden apreciar ciertas diferencias en la distribución de los descriptores tras aplicar t-SNE, lo cual sugiere que cada modalidad aporta información relevante y no redundante en el descriptor final, favoreciendo la robustez del método ante los errores de cada modalidad por separado.

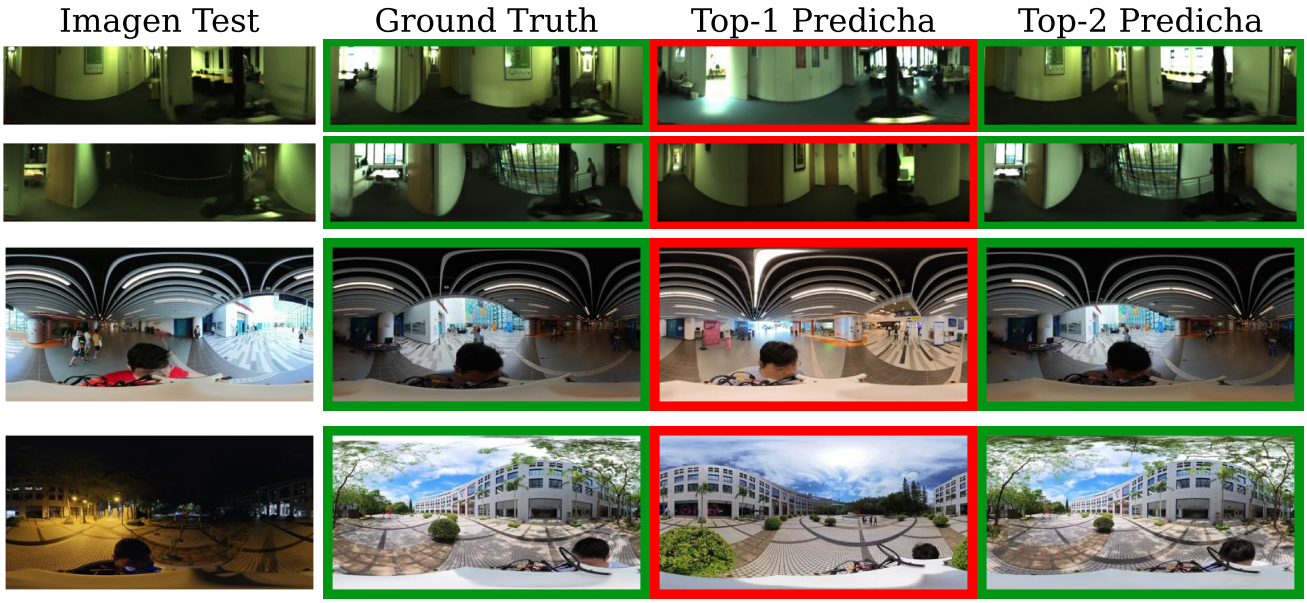


Figura 8: Ejemplos de situaciones de fallo del método propuesto. Los dos ejemplos superiores pertenecen a la base de datos COLD, mientras que los dos ejemplos inferiores pertenecen a la base de datos 360Loc. El borde de color rojo indica una predicción incorrecta, mientras que el borde de color verde indica una predicción correcta.

4.5.2. *Limitaciones*

Por último, se ha llevado a cabo un análisis de las limitaciones del método propuesto. Para ello, la Figura 8 muestra algunos casos en los cuales la mejor configuración del método (fusión tardía con $RGB + I + Mag$) no es capaz de predecir correctamente la posición donde se ha capturado la imagen de test.

A partir de los ejemplos de la Figura 8, se puede observar que el método propuesto puede fallar en situaciones especialmente desafiantes, tales como la ambigüedad perceptual o los cambios de iluminación extremos, especialmente en el último caso. No obstante, cabe destacar que para todos estos casos, la segunda predicción sí que se corresponde con la posición real de la imagen original dentro del mapa. Esto sugiere que, aunque el método no es totalmente inmune a situaciones extremas, la combinación de características multimodales preserva información suficiente para mantener la hipótesis correcta con un alto grado de confianza, tal y como muestran los resultados cuantitativos recogidos en la Sección 4.4, que confirman una mejoría significativa respecto a los métodos que emplean únicamente imágenes RGB. En trabajos futuros, se tratará de mejorar aún más la robustez de los métodos de VPR ante estas situaciones.

5. Conclusiones

La presente investigación evidencia que la integración de conocimiento previo estructural y fotométrico mediante características analíticas de bajo nivel mitiga considerablemente el déficit de generalización que sufren los modelos puros de aprendizaje profundo en la tarea de Reconocimiento Visual de Lugares (VPR). En lugar de depender exclusivamente de la información RGB, se concluye que la magnitud del gradiente y la intensidad aportan información complementaria que incrementa el rendimiento y la robustez del sistema frente a cambios de apariencia pronunciados, como variaciones de iluminación severas.

En relación con el método de fusión, el estudio concluye de forma consistente que la fusión tardía representa la estrategia óptima frente a las arquitecturas de fusión temprana o intermedia. Aislar las modalidades hasta la etapa de formación del descriptor permite explotar la capacidad de modelos preentrenados sin sufrir pérdidas de rendimiento por interferencias conjuntas, logrando mejoras significativas en la localización incluso sin la necesidad de aplicar técnicas de ajuste fino.

Desde la perspectiva de la literatura actual, este trabajo demuestra que es posible alcanzar una robustez comparable a la de arquitecturas pesadas basadas en Transformers Visuales, utilizando únicamente el procesamiento de imágenes capturadas por un único sensor omnidireccional. En consecuencia, este método se posiciona como una solución de VPR que optimiza el balance entre precisión computacional y eficiencia, siendo ideal para aplicaciones robóticas con recursos limitados.

Finalmente, los trabajos futuros se centrarán en desarrollar mecanismos de fusión dinámicos basados en atención por canales (*channel-attention*), así como en explorar la integración de esta arquitectura con dominios adicionales de información, tales como la semántica o la imagen térmica.

Agradecimientos

El Ministerio de Ciencia, Innovación y Universidades de España ha apoyado este trabajo a través de las “Ayudas para la Formación de Profesorado Universitario” (M. Alfaro, FPU23/00587; J.J. Cabrera, FPU21/04969). Este trabajo se enmarca dentro del proyecto PID2023-149575OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, EU. También forma parte del proyecto CIPROM/2024/8, financiado por la Generalitat Valenciana, Conselleria de Educación, Cultura, Universidades y Empleo (programa PROMETEO 2025).

Referencias

Alfaro, M., Cabrera, J. J., Reinoso, O., Gil, A., Payá, L., 2025. Localización visual mediante imágenes omnidireccionales y técnicas de fusión temprana. Jornadas de Automática (46).
DOI: https://doi.org/10.17979/ja-cea.2025.46.12239

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. En: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307.
DOI: https://doi.org/10.48550/arXiv.1511.07247

Berton, G., Masone, C., Caputo, B., June 2022. Rethinking visual geo-localization for large-scale applications. En: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4878–4888.
DOI: https://doi.org/10.48550/arXiv.2204.02287

Berton, G., Trivigno, G., Caputo, B., Masone, C., 2023. EigenPlaces: Training viewpoint robust models for visual place recognition. En: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11080–11090.
DOI: https://doi.org/10.48550/arXiv.2308.10832

Cabrera, J. J., Alfaro, M., Gil, A., Reinoso, O., Payá, L., 2026. Robust place recognition under illumination changes using pseudo-LiDAR from omnidirectional images. Scientific Reports 16 (8817).
DOI: https://doi.org/10.1038/s41598-026-39848-y

Cabrera, J. J., Santo, A., Gil, A., Viegas, C., Payá, L., 2025. MinkUNeXt: point cloud-based large-scale place recognition using 3D sparse convolutions. Array, 100569.
DOI: 10.1016/j.array.2025.100569

Chen, C., Wang, B., Lu, C. X., Trigoni, N., Markham, A., 2023. Deep learning for visual localization and mapping: A survey. IEEE Transactions on Neural Networks and Learning Systems 35 (12), 17000–17020.
DOI: https://doi.org/10.1109/TNNLS.2023.3309809

Chen, X., Chu, F.-J., Gleize, P., Liang, K. J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., *et al.*, 2025. Sam 3D: 3Dfy anything in images. arXiv preprint arXiv:2511.16624.
DOI: https://doi.org/10.48550/arXiv.2511.16624

Di Giammarino, L., Aloise, I., Stachniss, C., Grisetti, G., 2021. Visual Place Recognition using LiDAR intensity information. En: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 4382–4389.
DOI: https://doi.org/10.1109/IROS51168.2021.9636649

Fan, Z., Zhang, L., Wang, X., Shen, Y., Deng, F., 2025. LiDAR, IMU, and camera fusion for simultaneous localization and mapping: a systematic review. Artificial Intelligence Review 58 (6), 1–59.
DOI: https://doi.org/10.1007/s10462-025-11187-w

Flores, M., Valiente, D., Peidró, A., Reinoso, O., Payá, L., 2024. Generating a full spherical view by modeling the relation between two fisheye images. The Visual Computer 40 (10), 7107–7132.
DOI: 10.1007/s00371-024-03293-7

Gálvez-López, D., Tardos, J. D., 2012. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on robotics 28 (5), 1188–1197.
DOI: https://doi.org/10.1109/TRO.2012.2197158

Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., Soda, P., 2025. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image and Vision Computing 158, 105509.
DOI: https://doi.org/10.1016/j.imavis.2025.105509

Hausler, S., Jacobson, A., Milford, M., 2019. Multi-process fusion: Visual place recognition using multiple image processing methods. IEEE Robotics and Automation Letters 4 (2), 1924–1931.
DOI: https://doi.org/10.1109/LRA.2019.2898427

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. En: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.
DOI: https://doi.org/10.48550/arXiv.1512.03385

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737.
DOI: https://doi.org/10.48550/arXiv.1703.07737

Huang, H., Liu, C., Zhu, Y., Cheng, H., Braud, T., Yeung, S.-K., June 2024. 360Loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. En: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22314–22324.
DOI: 10.48550/arXiv.2311.17389

Izquierdo, S., Civera, J., 2024. Optimal transport aggregation for Visual Place Recognition. En: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 17658–17668.
DOI: https://doi.org/10.48550/arXiv.2311.15937

Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., Garg, S., 2023. AnyLoc: Towards universal visual place recognition. IEEE Robotics and Automation Letters 9 (2), 1286–1293.
DOI: https://doi.org/10.1109/LRA.2023.3343602

Lai, H., Yin, P., Scherer, S., 2022. Adafusion: Visual-LiDAR fusion with adaptive weights for place recognition. IEEE Robotics and Automation Letters 7 (4), 12038–12045.
DOI: https://doi.org/10.1109/LRA.2022.3210880

Li, Z., Shang, T., Xu, P., Deng, Z., 2025. Place recognition meet multiple modalities: a comprehensive review, current challenges and future development. Artificial Intelligence Review 58 (11), 363.
DOI: https://doi.org/10.1007/s10462-025-11367-8

Lin, H., Chen, S., Liew, J., Chen, D. Y., Li, Z., Shi, G., Feng, J., Kang, B., 2025. Depth Anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647.
DOI: https://doi.org/10.48550/arXiv.2511.10647

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., June 2022. A ConvNet for the 2020s. En: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986.
DOI: https://doi.org/10.48550/arXiv.2201.03545

Maaten, L. v. d., Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research 9 (Nov), 2579–2605.
DOI: 10.5555/3045118.3045270

Máximo, M., Santo, A., Gil, A., Ballesta, M., Valiente, D., 2026. A coarse-to-fine 3D LiDAR localization with deep local features for long-term robot navigation in large environments. International Journal of Intelligent Systems 2026 (1), 4278222.
DOI: 10.1155/int/4278222

Menegatti, E., Maeda, T., Ishiguro, H., 2004. Image-based memory for robot navigation using properties of omnidirectional images. Robotics and Autonomous Systems 47 (4), 251–267.
DOI: https://doi.org/10.1016/j.robot.2004.03.014

Payá, L., Reinoso, O., Berenguer, Y., Úbeda, D., 2016. Using omnidirectional vision to create a model of the environment: A comparative evaluation of global-appearance descriptors. Journal of Sensors 2016 (1), 1209507.
DOI: https://doi.org/10.1155/2016/1209507

Pronobis, A., Caputo, B., 2009. COLD: The CoSy localization database. The International Journal of Robotics Research 28 (5), 588–594.
DOI: 10.1177/0278364909103912

Rostkowska, M., Skrzypczyński, P., 2023. Optimizing appearance-based localization with catadioptric cameras: small-footprint models for real-time inference on edge devices. Sensors 23 (14), 6485.
DOI: https://doi.org/10.3390/s23146485

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., June 2020. Superglue: Learning feature matching with graph neural networks. En: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4938–4947.
DOI: https://doi.org/10.48550/arXiv.1911.11763

Siagian, C., Itti, L., 2007. Rapid biologically-inspired scene classification using features shared with visual attention. IEEE transactions on pattern analysis and machine intelligence 29 (2), 300–312.
DOI: https://doi.org/10.1109/TPAMI.2007.40

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
DOI: https://doi.org/10.48550/arXiv.1409.1556

Wang, T.-H., Huang, H.-J., Lin, J.-T., Hu, C.-W., Zeng, K.-H., Sun, M., 2018. Omnidirectional CNN for visual place recognition and navigation. En: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 2341–2348.
DOI: https://doi.org/10.1109/ICRA.2018.8463173

Zhou, Z., Xu, J., Xiong, G., Ma, J., 2023. LCPR: A multi-scale attention-based LiDAR-camera fusion network for place recognition. IEEE Robotics and Automation Letters 9 (2), 1342–1349.
DOI: https://doi.org/10.1109/LRA.2023.3346753

Zhu, Q., Yeh, M.-C., Cheng, K.-T., Avidan, S., 2006. Fast human detection using a cascade of histograms of oriented gradients. En: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). Vol. 2. IEEE, pp. 1491–1498.
DOI: https://doi.org/10.1109/CVPR.2006.119