ICINCO 2025

22nd International Conference on Informatics in Control, Automation and Robotics

PROCEEDINGS

Volume 1

Marbella, Spain

20 - 22 October, 2025

EDITORS

Giuseppina Carla Gini
Radu-Emil Precup
Dimitar Filev

https://icinco.scitevents.org

SPONSORED BY

INSTICC

PAPERS AVAILABLE AT

SCITEPRESS DIGITAL LIBRARY

ICINCO 2025

Proceedings of the
22nd International Conference on
Informatics in Control, Automation and Robotics

Volume 1

Marbella - Spain

October 20 - 22, 2025

Sponsored by
INSTICC - Institute for Systems and Technologies of Information, Control and Communication

IEEE Technically co-sponsored by
IEEE SMC - TC on Evolving Intelligent Systems

Technically Co-sponsored by
IFAC - International Federation of Automatic Control

ACM In Cooperation
SIGAI - ACM Special Interest Group on Artificial Intelligence

In Cooperation with
AAAI - Association for the Advancement of Artificial Intelligence
INNS - International Neural Network Society

Edited by Giuseppina Gini, Radu-Emil Precup and Dimitar Filev

Printed in Portugal
ISSN: 2184-2809
ISBN: 978-989-758-770-2
DOI: 10.5220/0000205300003982
Depósito Legal: 551123/25

https://icinco.scitevents.org
icinco.secretariat@insticc.org

CONTENTS

Place Recognition with Omnidirectional Imaging and Confidence-Based Late Fusion

Marcos Alfaro[1][a], Juan José Cabrera[1][b], Enrique Heredia[1][c], Oscar Reinoso[1,2][d], Arturo Gil[1][e] and Luis Paya[1,2][f]

[1]*Research Institute for Engineering (I3E), Miguel Hernández University, Elche, Spain*

[2]*Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain*

{*malfaro, juan.cabreram, e.heredia, o.reinoso, arturo.gil, lpaya*}*@umh.es*

Keywords: Mobile Robotics, Place Recognition, Omnidirectional Cameras, Deep Learning, Late Fusion.

Abstract: Place recognition is crucial for the safe navigation of mobile robots. Vision sensors are an effective solution to address this task due to their versatility and low cost, but the images are sensitive to changes in environmental conditions. Multi-modal approaches can overcome this limitation, but the integration of different sensors often leads to larger computing and hardware costs. Consequently, this paper proposes enhancing omnidirectional views with additional features derived from them. First, feature maps are extracted from the original omnidirectional images. Second, each feature map is processed by an independent deep network and embedded into a descriptor. Finally, embeddings are merged by means of a late approach that weights each feature according to the confidence in the prediction of the networks. The experiments conducted in indoor and outdoor scenarios revealed that the proposed method consistently improves the performance across different environments and lighting conditions, presenting itself as a precise, cost-effective solution for place recognition. The code is available at the project website: https://github.com/MarcosAlfaro/VPR_LF_VisualFeatures.

1 INTRODUCTION

Robust and reliable localization is a cornerstone of autonomous systems, enabling mobile robots and autonomous vehicles to navigate complex environments safely and efficiently (Liu et al., 2024). In this context, Visual Place Recognition (VPR) consists in identifying the current location of a robot by matching the view captured by an onboard camera against a pre-existing map of visual landmarks. The recent success of deep learning, particularly with architectures like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has led to significant advancements in learning discriminative image descriptors for this task (Arandjelovic et al., 2016; Dosovitskiy et al., 2020). While these methods demonstrate high accuracy, they often exhibit a tendency to overfit, leading to a narrow window of optimal performance.

Among vision sensors, omnidirectional cameras are particularly advantageous for VPR. Their large field of view (up to 360°) provides comprehensive information about the robot's surroundings, offering a degree of inherent invariance to the robot's orientation (Cabrera et al., 2021).

However, the use of omnidirectional images entails significant challenges. First, these images suffer from severe geometric distortions, which can degrade the performance of deep learning models typically trained with regular pin-hole (perspective) images. Second, like all vision-based methods, omnidirectional VPR is highly susceptible to environmental appearance changes caused by variations in lighting, weather, and seasons, as well as perceptual aliasing, where distinct locations appear visually similar.

A common strategy to overcome the limitations of a single sensor modality is to fuse its data with information from other sensors, such as LiDAR. This multi-modal approach leverages the strengths of each sensor, for example, by combining the rich appearance information from a camera with the geometric precision of LiDAR to build more robust environmental representations (Yu et al., 2022). However, the integration of multiple sensor types increases the hard-

[a] https://orcid.org/0009-0008-8213-557X
[b] https://orcid.org/0000-0002-7141-7802
[c] https://orcid.org/0009-0001-7717-1428
[d] https://orcid.org/0000-0002-1065-8944
[e] https://orcid.org/0000-0001-7811-8955
[f] https://orcid.org/0000-0002-3045-4316

Alfaro, M., Cabrera, J. J., Heredia, E., Reinoso, O., Gil, A. and Paya, L.
Place Recognition with Omnidirectional Imaging and Confidence-Based Late Fusion.

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 1*, pages 117-125
ISBN: 978-989-758-770-2; ISSN: 2184-2809

ware and computational cost and system complexity, which can be prohibitive for many robotic platforms.

This paper introduces an alternative paradigm: instead of adding new sensor modalities, we propose to use only one type of sensor (omnidirectional camera) and enhance this visual information with additional features. These features, which are less sensitive to photometric variations than standard RGB channels, are treated as independent information streams. They are then integrated with the original image data using a novel, adaptive late fusion strategy. This fusion mechanism operates as a weighted sum, where the contribution of each feature stream is dynamically determined by its confidence score during the retrieval process. This allows the system to intelligently rely on the most discriminative representation for any given query.

Consequently, the proposed method adds the robustness of hand-crafted features to the high accuracy and efficiency of CNNs for VPR, aiming for a both accurate and robust solution for this task. Therefore, the primary contributions of this work are twofold:

- We leverage some features derived from each original image to increase the robustness against challenging illumination and appearance variations that are common in real-world environments.
- We introduce a novel fusion technique that dynamically weights each feature stream based on its retrieval confidence. This allows the model to adaptively prioritize the most reliable information source, significantly improving VPR accuracy under challenging conditions.

The remainder of this manuscript is structured as follows. Section 2 reviews the state of the art. In Section 3, the proposed method is detailed. Section 4 describes the experiments. Finally, conclusions and future work are discussed in Section 5.

2 RELATED WORK

2.1 Visual Place Recognition

The role of VPR is crucial for the safe localization and navigation of mobile robots, and extensive research has been performed in the design of new models and techniques to address this task (Schubert et al., 2023). Early approaches relied on hand-crafted features to create global image descriptors but, with the rise of artificial intelligence, deep networks are widely employed currently as image encoders (Arandjelovic et al., 2016; Oquab et al., 2023).

CNN-based models, such as CosPlace (Berton et al., 2022) and EigenPlaces (Berton et al., 2023), offer remarkable efficiency and accuracy. More recently, ViTs have emerged as a powerful alternative, demonstrating exceptional performance due to their ability to capture global context. However, their complexity and data requirements often necessitate sophisticated training strategies, such as the use of adapters. Notable ViT-based models include AnyLoc (Keetha et al., 2023), SALAD (Izquierdo and Civera, 2024) and SelaVPR (Lu et al., 2024). Concurrently, innovations in feature aggregation have further pushed the performance boundaries of both CNNs and ViTs (Ali-Bey et al., 2023).

2.2 Data Fusion

In some occasions, mobile robots are equipped with multiple exteroceptive sensors, whose data are fused to reduce uncertainty and increase accuracy at VPR. For instance, vision sensors are frequently combined with LiDAR (Komorowski et al., 2021) and depth (Finman et al., 2015), among others.

Concerning the stage in which feature fusion is performed, these strategies are broadly divided into early fusion, in which sensory information is fused before being processed by a deep network (Heredia-Aguado et al., 2025), middle fusion, in which sensory data interact throughout the different layers of the model (Liu et al., 2022), and late fusion, where the different modalities are processed independently and their respective descriptors are subsequently fused (Komorowski et al., 2021).

Late fusion is commonly addressed with classical methods, such as concatenation or addition, as they achieve fairly competitive results in VPR. Also, there are end-to-end methods which fuse embeddings from different modalities through MLPs, but they usually show lower performance than previous methods (Komorowski et al., 2021). Besides, weighted sum is a suitable option, but the fusion weights are often non-dynamic and set empirically.

In this paper, a framework to enhance the visual data with intrinsic features derived from the original omnidirectional images is proposed. The visual data and these intrinsic features are merged by means of a late fusion approach that consists in a weighted sum, where the fusion weights are dynamic and calculated considering the confidence in the prediction of the trained models. The aim of this method is to improve the performance and robustness in VPR against challenging conditions while preserving a lightweight and cost-effective solution.

3 METHODOLOGY

3.1 Omnidirectional Vision

Omnidirectional cameras are characterized by their wide field of view, which enables the generation of comprehensive image descriptors that are inherently more robust to viewpoint variations. In this paper, two distinct types of omnidirectional vision systems are utilized to address VPR:

- Catadioptric system: This system comprises a standard camera paired with a hyperbolic mirror. It operates by capturing light rays that reflect off the mirror's surface towards the mirror's focus, where the camera's optical center is positioned. For our experiments, the resulting images were unwarped into a panoramic format.
- 360° camera: This type of sensor captures a full spherical image, providing a 360° field of view on all axes. A key advantage of modern 360° cameras is their capacity for high-resolution imaging. The spherical images captured by this camera are processed using an equirectangular projection to generate panoramic views.

3.2 Visual Features

While raw visual data are valuable for robotic scene understanding, their reliability can be compromised by challenges such as appearance variations (e.g., due to lighting or seasonal shifts) and visual aliasing. To mitigate these issues, we enhance the visual information by extracting a set of fundamental features from the omnidirectional images. These features, described below, provide alternative representations of the scene. Figures 1 and 2 display examples of the feature maps generated from sample images $Im(R,G,B)$.

- Intensity: Represents the brightness of each pixel and is calculated as the average value of blue (B), green (G) and red (R) color channels:

$$I = \frac{R+G+B}{3}. \tag{1}$$

- Hue: Represents the pure color component of each pixel. It is defined by the equation:

$$H = \cos^{-1}\left[\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}\right]. \tag{2}$$

- Gradient: Represents the intensity change in the local neighborhood of a pixel. The gradient is computed using Sobel operators, which are represented by the following convolution kernels for the vertical and horizontal axes, respectively:

$$g_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0^* & 0 \\ 1 & 2 & 1 \end{bmatrix}; \; g_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0^* & 2 \\ -1 & 0 & 1 \end{bmatrix}. \tag{3}$$

From the Sobel responses $\Delta_x = g_x(Im)$ and $\Delta_y = g_y(Im)$, two distinct features are derived:

– Magnitude: The gradient magnitude is calculated as the sum of the absolute intensity variations along both axes:

$$Mag = |\Delta x| + |\Delta y|. \tag{4}$$

– Orientation: The gradient orientation is defined as the direction of the maximum intensity variation and is given by:

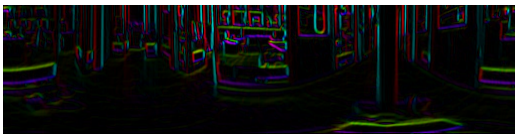
$$\theta = \arctan 2\,(\Delta y, \Delta x). \tag{5}$$

(a) Original

(b) Intensity

(c) Hue

(d) Gradient (Magnitude)



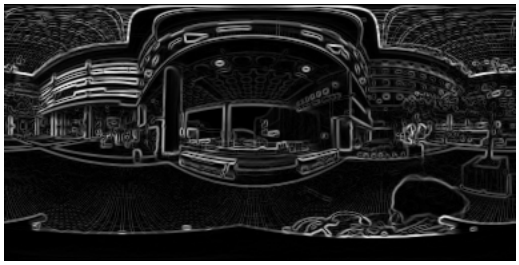
(e) Gradient (Orientation)

Figure 1: (a) Example of a panoramic image from the COLD database (Pronobis and Caputo, 2009) and feature maps obtained from the image: (b) intensity, (c) hue, (d) gradient magnitude and (e) gradient orientation.

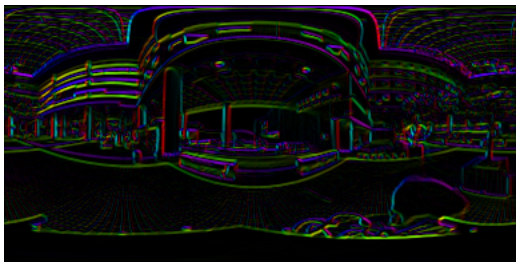
(a) Original

(b) Intensity

(c) Hue



(d) Gradient (Magnitude)



(e) Gradient (Orientation)

Figure 2: (a) Example of a panoramic image from the 360Loc database (Huang et al., 2024) and feature maps obtained from the image: (b) intensity, (c) hue, (d) gradient magnitude and (e) gradient orientation.

3.3 Data Fusion Approach

To create a more robust descriptor, information from the raw images and the extracted features is combined using a late fusion strategy. In this approach, each input stream (e.g., RGB, Hue, etc.) is processed by an independent neural network model to generate a feature-specific descriptor. These individual descriptors are then merged into a single, unified descriptor through a dynamic weighted sum, as shown in the following equation:

$$\vec{d}^{q} = \omega_{RGB} * \vec{d}_{RGB} + \omega_{I} * \vec{d}_{I} + \\ + \omega_{Hue} * \vec{d}_{Hue} + \omega_{Mag} * \vec{d}_{Mag} + \omega_{\theta} * \vec{d}_{\theta}, \tag{6}$$

where $\vec{d}_i$ represents the descriptor for each feature type and ω_i is its corresponding weight. These weights are calculated as detailed in Section 3.5.

3.4 Model Selection and Adaptation

To generate global descriptors from the omnidirectional images and their feature maps, we employed CosPlace (Berton et al., 2022), a state-of-the-art CNN pre-trained on 41.2 millions of images for the VPR task. From the available architectures, a comparative evaluation is conducted in Section 4.3.1 to select the optimal backbone for each database.

We adopted a transfer learning strategy, adapting the pre-trained model to process our specific input types. For the single-channel feature maps, the model's input layer, originally designed for 3-channel RGB images, was modified to accept a single-channel input. The initial weights for this modified layer were set by averaging the pre-trained weights of the original R, G, and B input channels. The entire network was then fine-tuned for each specific feature stream.

3.5 Training and Evaluation

During the training stage, an independent CosPlace model was fine-tuned for each of the five input streams: raw RGB images and the four derived visual feature maps. A triplet architecture was employed, which involves training the network with triplets of images: an anchor (I_a), a positive (I_p) and a negative (I_n) sample, chosen in such a way that the distance between the capture points of the anchor and the positive images must be lower than a threshold distance r_p, and the anchor and negative images must be captured further apart than a threshold distance r_n, being $r_p <= r_n$. The objective is to train the network to produce similar descriptors for images captured from

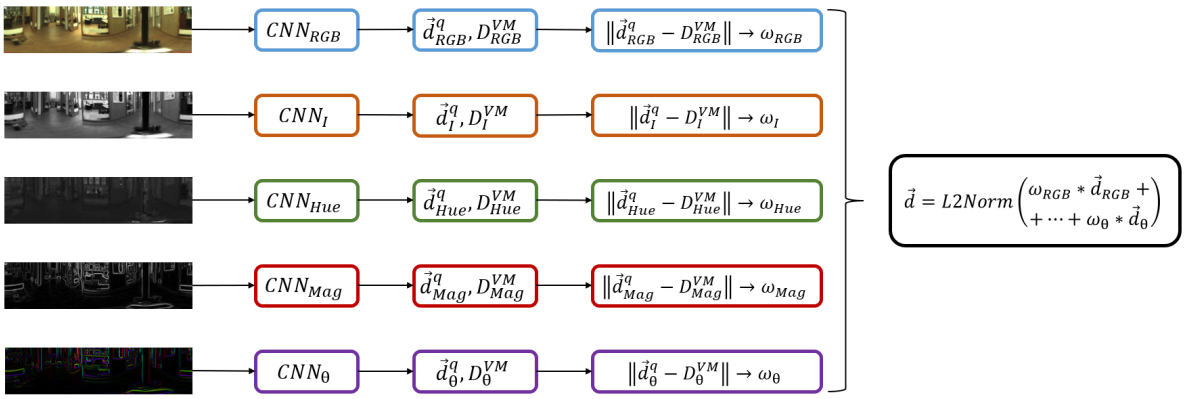


Figure 3: General outline of the proposed late fusion method.

close positions and dissimilar descriptors for images of different places.

For evaluation, a global descriptor $\vec{d}^q$ is generated for each image I_{query} using the late fusion method outlined in Figure 3. The process is as follows:

1. For each feature stream k, the query image is passed through its corresponding network (net_k) to produce a descriptor, $\vec{d}_k$.
2. This descriptor $\vec{d}_k$ is compared against all the descriptors in the database for that feature, $D^{VM}_k = \left[\vec{d}_{k_1}, ..., \vec{d}_{k_n}\right]$, using the Euclidean distance.
3. A confidence score, which serves as the fusion weight ω_k, is calculated. This weight rewards feature streams that produce a highly confident match (i.e., the best-matching descriptor is significantly closer than other descriptors in the database). The weight is given by:

$$\omega_k = \frac{\frac{1}{dist_{min}}}{\sum_{i=1}^{n} \frac{1}{dist_i}}, \tag{7}$$

   where $dist_{min}$ is the smallest distance found in Step 2, $dist_i$ is the distance between the query descriptor and the i-th descriptor from the visual model and n is the number of images from the visual model.
4. After repeating this process for all five feature streams, the final fused query descriptor, $\vec{d}^q$, is calculated as the weighted sum of the individual query descriptors using Equation 6.

Once the descriptor $\vec{d}^q$ is generated, it is compared with the visual map $D^{VM} = \left[\vec{d}_1, ..., \vec{d}_n\right]$, where each entry is a pre-computed descriptor also created as the weighted average of the five feature descriptors for that location. The minimum Euclidean distance indicates the retrieved position in the map $I_r = (x_r, y_r)$.

The retrieval error, e^j_r, for a given query j, is the geometric distance between the ground-truth position of the query image and the retrieved position.

To quantify the performance of our method, the Recall@1 ($R@1$) metric is used. This measures the percentage of query images that are correctly localized within a specified threshold distance, d:

$$R@1(\%) = \frac{\sum_{j=1}^{M} \mathbb{I}(e^j_r \leq d)}{M} \times 100, \tag{8}$$

where M is the number of images in the test set and $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the condition is true and 0 otherwise.

4 EXPERIMENTS

4.1 Datasets

To evaluate the proposed method under challenging conditions, the experiments were conducted on two distinct datasets: COLD and 360Loc, representing an indoor environment and a mixed indoor-outdoor scenario, respectively.

4.1.1 COLD

The COLD database (Pronobis and Caputo, 2009) consists of panoramic images captured with a catadioptric camera system across several indoor environments: Freiburg Part A (FR-A) and B (FR-B), and Saarbrücken Part A (SA-A) and B (SA-B). To evaluate the robustness to appearance changes, images were captured under three different lighting conditions: cloudy, night, and sunny. Table 1 details the number of images used for the training and test sets.

Table 1: Image sets employed for training and evaluation from the COLD database. *Training set.

Environment	Train/ Database	Test Cloudy	Test Night	Test Sunny
FR-A	556*	2595	2707	2114
FR-B	560	2008	-	1797
SA-A	586	2774	2267	-
SA-B	321	836	870	872

4.1.2 360Loc

The 360Loc dataset (Huang et al., 2024) contains high-resolution, equirectangular images captured in four distinct semi-open locations: atrium, concourse, hall, and piatrium. The images were collected under day and night conditions, making the dataset suitable for evaluating performance under severe lighting variations. Table 2 shows the number of images in the training and test sets.

Table 2: Image sets employed for training and evaluation from the 360Loc database. *Training set.

Environment	Train / Database	Test Day	Test Night
atrium	581*	875	1219
concourse	491	593	514
hall	540	1123	1061
piatrium	632	1008	697

4.2 Implementation Settings

To conduct the experiments, the Lazy Triplet Loss (Uy and Lee, 2018) was employed, with a margin $m = 0.5$ and a batch size $N = 4$, as it provided great localization accuracy in similar works (Komorowski et al., 2021). The selected optimizer algorithm was the SGD (Stochastic Gradient Descent) with a learning rate $lr = 0.001$. All the experiments were performed on an NVIDIA GeForce RTX 4080 SUPER GPU with 16GB of memory.

Concerning the triplet sample selection during the training process (see Section 3.5), r_p and r_n were both set to 0.4m for the COLD database. For the 360Loc database, r_p and r_n were set to 2m and 5m, respectively. These values were chosen to perform a training with challenging and varied samples, considering the number of training images and the dimensions of the environments. Regarding the threshold distance d to calculate R@1, d was set to 0.5m for the COLD database, $5m$ for the concourse environment from the 360Loc dataset, and 10m for the rest of the environments of 360Loc, according to the criteria followed in (Pronobis and Caputo, 2009) and (Huang et al., 2024).

4.3 Ablation Study

4.3.1 Backbone Selection

First, a preliminary experiment was conducted to select the optimal CNN backbone to embed the images and visual features into descriptors. All available CosPlace models, i.e. VGG16, ResNet-18, ResNet-50 and ResNet-101 were tested without training, with a descriptor size of 512. Figure 4 displays the R@1 results obtained with each backbone on both the COLD and 360Loc environments.

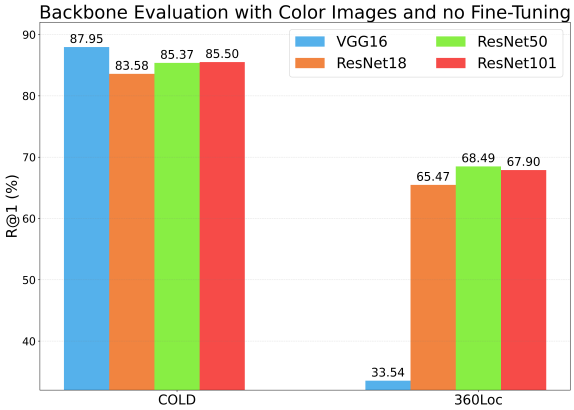


Figure 4: Backbone evaluation on both datasets.

As shown in Figure 4, since VGG16 and ResNet50 produce the best results the best results for indoor and outdoor experiments, respectively, they have been employed in the subsequent experiments.

4.3.2 Feature Evaluation Before Fusion

Next, each visual feature is evaluated independently. For this purpose, a separate model was fine-tuned and tested for each of the five feature streams on both the COLD and 360Loc datasets. The overall Recall@1 (R@1) for each feature is presented in Figure 5.

As shown in Figure 5, the model trained on the original RGB images (baseline) achieved the highest performance. This result is expected, as the CosPlace model was pre-trained on standard color images. Nonetheless, the models trained on the intensity and gradient magnitude features demonstrated competitive performance across both datasets, validating their potential as robust modalities for place recognition.

Besides, Figure 6 shows the average confidence in the predictions of these models in the COLD Freiburg environment. These results are separated into correct and incorrect retrievals.

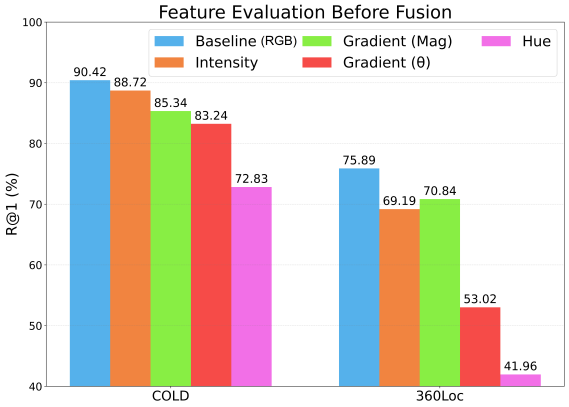


Figure 5: R@1 obtained by models trained with each visual feature in both datasets.

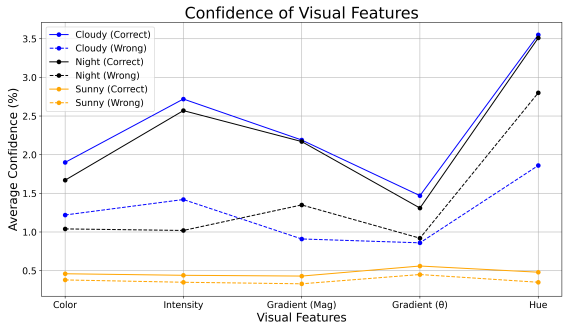


Figure 6: Average confidence of models trained with each visual feature in the COLD Freiburg A environment.

From Figure 6, it can be observed that, for every feature, the model exhibits higher confidence when it makes a correct retrieval. This supports the use of the confidence to build descriptors that combine different visual features. It can also be noticed that the model trained with hue shows the highest confidence compared to the rest of features, even in wrong predictions. For this reason, besides its fairly low R@1, hue is not a suitable feature for the proposed method.

4.3.3 Feature Evaluation after Fusion

Next, we evaluated the proposed late fusion approach by combining the baseline RGB model with the other visual feature streams. Several combinations were tested to identify the most effective fusion strategy across different scenarios and lighting conditions. Tables 3 and 4 present the detailed R@1 scores on the COLD and 360Loc datasets, respectively.

On the indoor COLD dataset (Table 3), the best global performance was achieved by employing the baseline RGB model along with the intensity feature. On the mixed-environment 360Loc dataset (Table 4), combining RGB with both intensity and gradient magnitude ($I + Mag$) produced the largest improvement, achieving a global R@1 of 80.29%, a +4.40% increase in performance compared to the baseline.

Notably, the combination of intensity and gradient magnitude also demonstrated highly competitive performance under the most challenging lighting conditions, such as sunny for the indoor dataset and night for the outdoor dataset, where traditional color-based methods often struggle.

To better understand how different features contribute to localization, Figure 7 displays the feature that dominated the fusion process for different query images. On these maps, each point marks the location of a query image. The color indicates which feature yielded the highest confidence (ω_k) for that query. A dot · signifies a successful localization (retrieval error $\leq d$), while a cross ($\times$) denotes a failure.

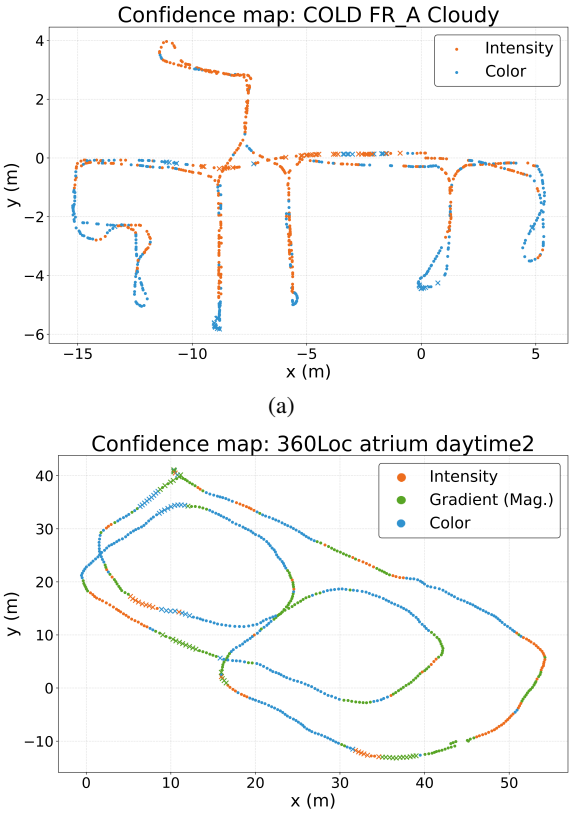


Figure 7: Confidence maps for (a) COLD (FR-A cloudy) and (b) 360Loc (atrium daytime).

4.3.4 Comparison of Late Fusion Methods

Finally, we benchmarked our proposed confidence-based late fusion method against other conventional fusion techniques. For this comparison, we used the best-performing feature combination identified for

Table 3: R@1 after late fusion in the COLD database at every environment and lighting condition.

Features	FR-A			FR-B		SA-A		SA-B			Global
	Cloudy	Night	Sunny	Cloudy	Sunny	Cloudy	Night	Cloudy	Night	Sunny	
Baseline (RGB)	92.91	95.01	83.35	85.86	85.92	**76.74**	**64.31**	88.35	78.51	83.94	83.59
+Intensity (*I*)	**93.10**	95.27	83.82	85.31	89.59	**76.74**	63.21	**91.99**	78.74	84.75	**84.25**
+Gradient (*Mag*)	91.48	95.20	84.72	84.56	91.04	75.29	60.65	89.47	78.74	83.72	83.49
+Gradient (θ)	91.64	94.98	82.12	84.41	93.21	74.93	50.64	83.73	**80.00**	81.88	81.75
+*Hue*	91.52	94.42	74.55	**86.10**	73.34	74.67	35.33	88.88	72.53	81.08	77.24
+I + Mag	92.72	95.38	84.39	84.36	92.21	75.51	61.18	90.19	78.62	84.98	83.95
+All \wo Hue	92.18	**95.71**	**87.51**	85.61	**94.10**	75.77	59.15	88.52	78.05	84.98	84.16
+All	92.22	95.16	85.48	85.91	92.82	75.42	51.08	90.07	76.44	**85.67**	83.06

Table 4: R@1 after late fusion in the 360Loc database at every environment and lighting condition.

Features	atrium		concourse		hall		piatrium		Global
	Day	Night	Day	Night	Day	Night	Day	Night	
Baseline (RGB)	**94.62**	73.58	88.46	73.35	91.00	54.51	**85.99**	45.62	75.89
+Intensity (*I*)	93.85	68.18	89.02	72.76	91.52	59.12	84.81	42.75	75.25
+Gradient (*Mag*)	93.05	67.91	90.37	**79.96**	92.67	63.35	84.47	**47.92**	77.46
+Gradient (θ)	89.71	65.29	87.85	73.15	90.47	49.70	78.00	28.26	70.30
+*Hue*	86.87	65.55	80.22	35.99	90.38	39.49	77.20	36.58	64.03
+I + Mag	92.50	67.16	**90.88**	79.38	91.97	**69.75**	85.24	47.20	**80.29**
+All \wo Hue	91.30	70.23	90.21	78.60	92.45	67.32	83.17	42.75	77.00
+All	91.30	**74.73**	88.85	77.43	**93.21**	66.32	80.47	42.18	76.81

each dataset (*RGB*+*I* for COLD, and *RGB*+*I*+*Mag* for 360Loc).

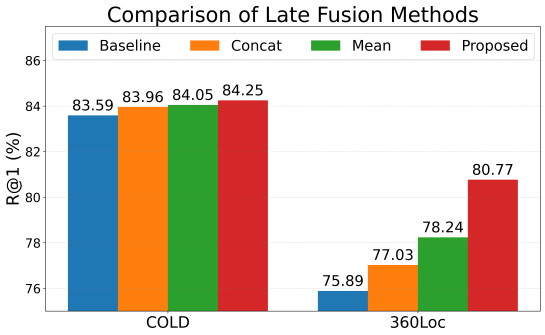


Figure 8: Comparison of late fusion methods in both datasets.

The results, presented in Figure 8, demonstrate that the proposed method consistently outperforms other techniques. The improvement is particularly pronounced in the challenging outdoor scenarios of 360Loc, highlighting the effectiveness of dynamically weighting feature streams based on model confidence.

5 CONCLUSIONS

In this manuscript, omnidirectional images are enriched with intrinsic visual features, such as the intensity and gradient magnitude, to tackle place recognition. These features are integrated through a confidence-based late fusion framework.

Our experimental results show that this approach consistently enhances VPR performance across diverse environments and lighting conditions. The performance gain is particularly significant in outdoor scenarios, which are prone to severe appearance changes. For indoor environments, fusing RGB with intensity information has yielded the best results, while a combination of intensity and gradient magnitude proves most effective for the mixed indoor-outdoor dataset. Crucially, our proposed dynamic fusion method demonstrates superior performance compared to conventional late fusion techniques.

Future works will focus on integrating additional data modalities such as estimated depth or semantic information, to further enrich the scene representation. Furthermore, we will study the use of attention mechanisms to conduct this data fusion.

ACKNOWLEDGEMENTS

The Ministry of Science, Innovation and Universities (Spain) has funded this work through FPU23/00587 (M. Alfaro) and FPU21/04969 (J.J. Cabrera). This work is part of the projects PID2023-149575OB-I00, funded by MICIU/AEI/10.13039/501100011033 and by FEDER UE, and CIPROM/2024/8, funded by Generalitat Valenciana.

REFERENCES

Ali-Bey, A., Chaib-Draa, B., and Giguere, P. (2023). MixVPR: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2998–3007.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.

Berton, G., Masone, C., and Caputo, B. (2022). Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888.

Berton, G., Trivigno, G., Caputo, B., and Masone, C. (2023). Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090.

Cabrera, J. J., Cebollada, S., Payá, L., Flores, M., and Reinoso, O. (2021). A robust CNN training approach to address hierarchical localization with omnidirectional images. In *ICINCO*, pages 301–310.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Finman, R., Paull, L., and Leonard, J. J. (2015). Toward object-based place recognition in dense rgb-d maps. In *ICRA Workshop Visual Place Recognition in Changing Environments, Seattle, WA*, volume 76, page 480.

Heredia-Aguado, E., Cabrera, J. J., Jiménez, L. M., Valiente, D., and Gil, A. (2025). Static early fusion techniques for visible and thermal images to enhance convolutional neural network detection: A performance analysis. *Remote Sensing*, 17(6).

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

Izquierdo, S. and Civera, J. (2024). Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17658–17668.

Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., and Garg, S. (2023). AnyLoc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9(2):1286–1293.

Komorowski, J., Wysoczańska, M., and Trzcinski, T. (2021). MinkLoc++: LiDAR and monocular image fusion for place recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Liu, W., Fei, J., and Zhu, Z. (2022). MFF-PR: Point cloud and image multi-modal feature fusion for place recognition. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 647–655. IEEE.

Liu, Y., Wang, S., Xie, Y., Xiong, T., and Wu, M. (2024). A review of sensing technologies for indoor autonomous mobile robots. *Sensors*, 24(4):1222.

Lu, F., Zhang, L., Lan, X., Dong, S., Wang, Y., and Yuan, C. (2024). Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

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

Schubert, S., Neubert, P., Garg, S., Milford, M., and Fischer, T. (2023). Visual place recognition: A tutorial [tutorial]. *IEEE Robotics & Automation Magazine*, 31(3):139–153.

Uy, M. A. and Lee, G. H. (2018). PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4470–4479.

Yu, X., Zhou, B., Chang, Z., Qian, K., and Fang, F. (2022). MMDF: Multi-modal deep feature based place recognition of mobile robots with applications on cross-scene navigation. *IEEE Robotics and Automation Letters*, 7(3):6742–6749.