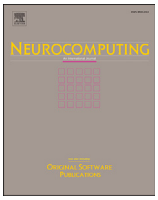
Neurocomputing 677 (2026) 133112

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



PDPR: Panoramic-depth place recognition through the fusion of visual and geometric-aware features

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

[a] Institute for Engineering Research, Miguel Hernández University, Avenida de la Universidad S/N, Elche, 03202, Comunidad Valenciana, Spain
[b] Valencian Graduate School and Research Network of Artificial Intelligence, Camí de Vera, Building 3Q, Valencia, 46020, Comunidad Valenciana, Spain

HIGHLIGHTS

- Monocular depth estimation is utilized to enhance place recognition.
- A thorough evaluation of preprocessing techniques is conducted to improve the depth maps.
- Fusion techniques are developed to leverage visual and geometric data.
- A model-agnostic approach that enhances the performance even without fine tuning.
- A robust method applicable across various scenarios and lighting conditions.

ARTICLE INFO

Communicated by D. Liu

Keywords:
Place recognition
Panoramic images
Monocular depth estimation
Data fusion

ABSTRACT

Omnidirectional cameras are a suitable and cost-effective choice for Visual Place Recognition (VPR), as they provide comprehensive information from the scene regardless of the robot orientation. However, vision sensors are vulnerable to environmental appearance changes (e.g., illumination, weather, season or moving objects). While multi-modal sensing approaches can overcome these challenges, they introduce significant cost and system complexity. This paper introduces PDPR (Panoramic-Depth Place Recognition), a novel fusion framework that enhances the robustness of VPR methods by integrating visual data with geometric features derived from monocular depth estimation techniques, while using a single-camera setup. In the ablation study, both early and late fusion strategies are evaluated to optimally combine appearance-based and depth-derived features. The extensive evaluation on challenging, indoor and outdoor datasets demonstrates that PDPR consistently boosts retrieval performance across multiple state-of-the-art VPR models. Furthermore, this improvement is achieved without requiring any fine tuning, allowing our method to function as a pluggable module for pretrained models. Consequently, this work presents a powerful, practical and low-cost solution for robust VPR, with high potential to scale as monocular depth estimation and VPR models continue to improve. The project website can be found at https://marcosalfaro.github.io/projects-PDPR/.

1. Introduction

Visual Place Recognition (VPR) aims to determine the location of a vehicle by matching a current query image against a database of previously visited places [1]. It has emerged as a cost-effective and scalable solution for applications that require localization or navigation capabilities in the field of mobile robotics. To address this task, images provide rich semantic and textural information at a low cost [2]. Omnidirectional cameras, in particular, are highly suitable for this task, as their 360° field of view (FoV) offers comprehensive scene information and inherent viewpoint invariance [3].

Despite these advantages, the reliance on visual appearance makes VPR algorithms vulnerable to environmental changes, such as different illumination conditions, seasons, or the presence of dynamic objects. While modern deep learning encoders, based on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have significantly advanced the capabilities in VPR [4,5], they still struggle when facing these pronounced appearance shifts.

* Corresponding author.
Email addresses: malfaro@umh.es (M. Alfaro), juan.cabreram@umh.es (J.J. Cabrera), arturo.gil@umh.es (A. Gil), o.reinoso@umh.es (O. Reinoso), lpaya@umh.es (L. Payá).

https://doi.org/10.1016/j.neucom.2026.133112
Received 19 November 2025; Received in revised form 27 January 2026; Accepted 18 February 2026
Available online 18 February 2026

To mitigate this, multi-modal approaches seek to improve robustness by fusing visual data with other sensory sources [6]. Information from LiDAR point clouds [7] or 3D depth sensors [8] is popular, as this geometric data is inherently robust to natural illumination changes (being self-illuminating sensors). However, this solution introduces significant system complexity, calibration overhead and cost. While RGB-D sensors [9] are an alternative, their limited field of view hinders the recognition of the complete scene, and their integration with the 360° FoV of omnidirectional camera setups is not straightforward.

Recently, Monocular Depth Estimation (MDE) has emerged as a powerful alternative, capable of generating dense, pixel-aligned depth maps from a single RGB image [10]. This technology provides a pathway to obtain rich geometric information about the scene without any additional sensing devices, perfectly complementing the existing camera. We hypothesize that fusing the geometric information from MDE-generated depth maps with visual data can significantly enhance VPR robustness against appearance changes, all while retaining a low-cost, single-camera setup. Building on this, we propose PDPR (Panoramic-Depth Place Recognition), a novel fusion framework that leverages state-of-the-art VPR models to encode panoramic images and their corresponding MDE-derived depth maps separately. These two embeddings —one capturing visual appearance, the other geometric structure— are then merged using an efficient fusion strategy.

As our experimental validation demonstrates, the proposed approach consistently improves retrieval performance across diverse environments, regardless of the employed VPR backbone. Importantly, this improvement is achieved without any model fine-tuning, allowing PDPR to act as a lightweight, pluggable module for existing, pre-trained systems. Therefore, the main contributions of this paper are:

- A novel pseudo-multimodal place recognition framework is proposed. It leverages both visual information from panoramic images and geometric data from MDE-generated depth maps, while using a low-cost and single-camera setup.
- A comprehensive evaluation of different fusion strategies (including early and late fusion) and depth preprocessing techniques is performed for optimal alignment between visual and depth features.
- The proposed solution is pluggable and model-agnostic. The experiments demonstrate that it increases the accuracy and robustness of existing VPR methods against pronounced changes in appearance, such as lighting variations, without requiring any fine tuning process.

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 and results. Finally, the conclusions and future work are discussed in Section 5.

2. State of the art

This section reviews the most recent approaches in Visual Place Recognition (Section 2.1), monocular depth estimation (Section 2.2) and multi-modal place recognition (Section 2.3).

2.1. Visual place recognition

Historically, Visual Place Recognition (VPR) was addressed with hand-crafted global-appearance descriptors, such as gist or HOG (Histogram of Oriented Gradients) [11], or with bags of words built with local descriptors [12]. With the rise of deep learning, these methods were replaced by a new generation of powerful image encoders. Convolutional Neural Networks (CNNs) like NetVLAD [13] and more recent Vision Transformer (ViT) architectures [14] have become the current standard for embedding images into compact, comprehensive embeddings.

Current research in VPR is focused on several key directions. One trend involves leveraging feature extraction backbones [15] from Vision Foundation Models (VFMs) like DINOv2 or Hiera [4,16,17], while another trend seeks to develop and train specific-task models for VPR [18], using large-scale datasets tailored for VPR [13,18–20]. Others seek to elaborate efficient training techniques to enhance robustness against changes in visual appearance or viewpoint [21]. Concurrently, new feature aggregation modules like MixVPR or SALAD [22,23] and hierarchical re-ranking pipelines [24] aim to improve both the efficiency and accuracy of the retrieval process.

Despite these significant advances, modern VPR methods remain highly sensitive to severe appearance shifts caused by illumination, weather and seasonal changes. Furthermore, the majority of these models are trained and benchmarked on standard pinhole camera images, overlooking the advantages of omnidirectional vision. While some work exists in this field [25,26], the design of robust, appearance-invariant solutions for this sensor modality remains a significant challenge.

2.2. Depth estimation

Monocular Depth Estimation (MDE), the task of predicting a dense depth map from a single RGB image, has seen remarkable progress in recent years. Since the introduction of the foundational MiDaS model [27], various more accurate architectures have been proposed [28–32]. In particular, Depth Anything v2 [10] has achieved state-of-the-art results across numerous MDE benchmarks. Concerning the specific geometry of 360° images, a parallel line of research has also emerged to develop MDE models specifically trained for panoramic and equirectangular images [33,34].

The high performance of MDE models has motivated their adoption in various downstream robotics tasks. For instance, they have been successfully used for depth completion [35], to enable 3D reconstruction from single images [36] and to serve as a geometric prior for Visual-SLAM systems [37]. This success in several related domains demonstrates that MDE can provide strong, yet largely unexplored, potential for enhancing VPR. Besides, other tasks like optical flow estimation benefit from geometric data for estimating motion from two images [38]. Furthermore, Han et al. [39] propose a transformer architecture with two different decoders, one for solving optical flow and the other for depth estimation.

2.3. Data fusion

To overcome the brittleness of purely visual methods, multi-modal place recognition has become a popular research area [40]. The core idea is to fuse visual information, which captures rich texture and appearance, with data from other types of sensors like LiDAR or RGB-D cameras [41,42]. The information provided by such sensors is inherently robust to the illumination and seasonal changes that typically cause VPR to fail.

The literature explores various fusion strategies, which are often categorized by the stage at which information is combined. Early fusion methods combine raw sensor data at the input level [43,44], middle fusion merges features within intermediate layers of a network [45,46] and late fusion combines the final 1-D descriptors from each modality [7,47]. Furthermore, related work in cross-modal place recognition has developed techniques to match data from different sensors, such as visual images to a LiDAR-based map [48,49].

3. Methodology

The primary objective of PDPR is to enhance the robustness of VPR by creating a comprehensive descriptor that captures both visual (appearance) and geometric (structural) information. The core idea is to leverage a Monocular Depth Estimation (MDE) model to generate a depth map from a single panoramic image, effectively creating a “pseudo-multi-modal” system from a single sensor. Our pipeline, which is summarized in Fig. 1, processes these two inputs —the RGB image and the depth map— to produce a final, fused embedding. This section details each component of the process.

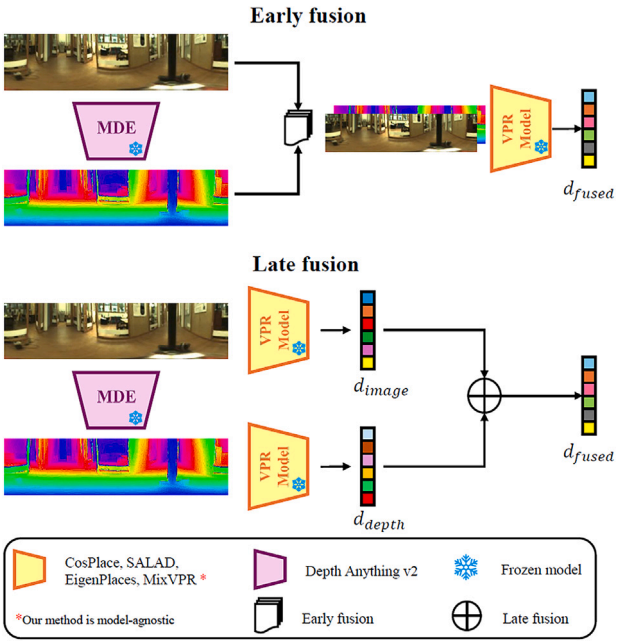


Fig. 1. General outline of the proposed method. First, depth maps are obtained by means of Depth Anything v2 [10]. Second, these depth maps are processed to adapt them to frozen VPR models, originally trained with standard, RGB images. Next, panoramic images and depth maps are combined through two different fusion approaches: early and late. The output is a global embedding that integrates both visual and depth data.

3.1. Depth anything v2

Depth Anything v2 (DAv2) [10] is employed in the current approach as a model to extract geometric information from images. The architecture of this state-of-the-art MDE model combines a DINOv2 encoder [17] with a Dense Prediction Transformer (DPT) decoder [50], and achieves outstanding performance across most MDE benchmarks.

The pre-trained DAv2-large model is employed, without any fine tuning, to generate relative depth maps from the panoramic RGB images. The output is a single-channel, 8-bit image (0-255). It is important to note that the output of DAv2 is inversely proportional to depth: nearby objects are represented by high-intensity values, while distant objects take low values. Examples of these generated depth maps are shown in Fig. 2.

3.2. Depth maps preprocessing

Current VPR models are commonly trained with 3-channel RGB images as input. In the current approach, we need to feed the model with a raw, single-channel depth map. This situation typically produces low results in terms of VPR. In this section, we propose and test different variations to adapt the depth data to the expectations of the VPR encoder. The proposed techniques are described below and an example for each case is shown in Fig. 3:

- Raw depth map (Fig. 3(b)): The 8-bit output from DAv2 is used directly.
- Histogram equalization (Fig. 3(c)): This contrast-enhancement technique is applied to prevent high-intensity (very close) objects from dominating the dynamic range of the depth map.
- Depth map inversion (Fig. 3(d)): The map is inverted ($D'(X,Y) = 255 - D(X,Y)$) so that high pixel values correspond to greater depth.
- Sharpening (Fig. 3(e)): A 3×3 sharpening kernel is applied to enhance the edges of objects, emphasizing the structural outlines of the scene.
- False color map (Fig. 3(f)): The single-channel depth map is projected into a 3-channel representation using a color map, e.g., HSV (Hue-Saturation-Value). This creates a 3-channel image that structurally resembles an RGB input, making it highly compatible with the pre-trained VPR encoder.

3.3. VPR backbone encoder

A key hypothesis of this work is that geometric information can boost pre-trained VPR models without costly retraining or fine-tuning processes, showing a model-agnostic performance improvement, which is crucial for practical applications (this property is assessed in Section 4.4.3). However, to ensure consistency during the ablation study, the CosPlace model [18] is selected for its strong performance and efficient architecture. In the experimental section, the proposed approach is evaluated in two different ways: by keeping the model frozen (i.e., the weights of the pre-trained model are not modified while obtaining the image and depth embeddings) and also by performing a fine tuning (i.e., the weights are modified). Among the available backbones, VGG-16 [51] is employed, with a descriptor size of 512.

3.4. Data fusion approaches

In this paper, we posit that the geometric information that is present in depth maps can be a suitable addition to color, and textural data from standard images. Throughout the experimental section, two different families of fusion methods are proposed, which are depicted in Fig. 1: early fusion (combining inputs before the encoder) and late fusion (combining descriptors after the encoder).

3.4.1. Early fusion

In this approach, the RGB image and the preprocessed depth map are merged at the channel level before being fed to the single VPR encoder. This requires modifying the input layer of the encoder to accept more than three channels (except for 3-channel variants which are described next).

Let R_c, G_c, B_c be the channels of the color image, D_d be the single-channel preprocessed depth map (i.e., sharpened), H_d, S_d, V_d be the

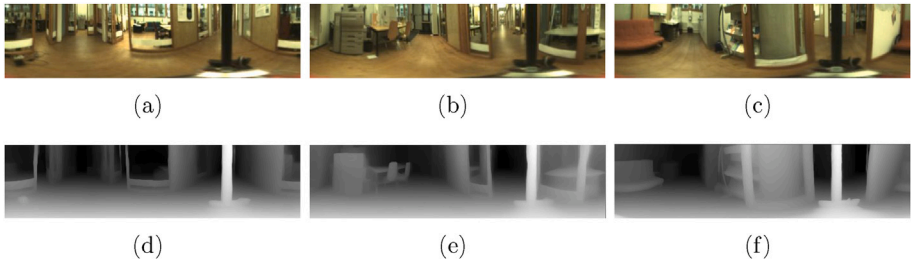


Fig. 2. Examples of (a,b,c) panoramic images captured under different lighting conditions and (d,e,f) their corresponding depth maps obtained with Depth Anything v2 [10].

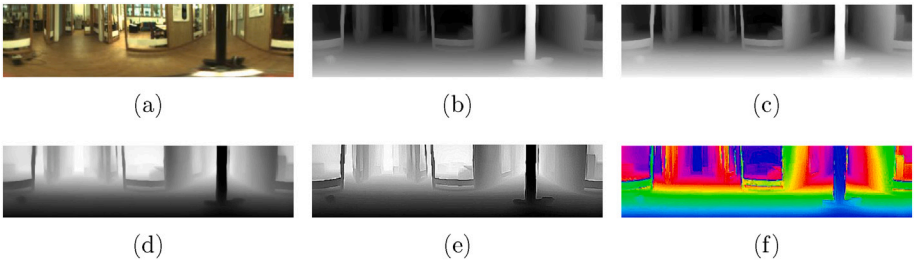


Fig. 3. Examples of (a) a panoramic image from the COLD database, (b) a raw depth map obtained by Depth Anything v2 (no preprocessing), (c) the depth map after histogram equalization, (d) inverse, (e) sharpening and (f) application of HSV colormap.

channels of the false-color (HSV) depth map and let || denote channel-wise concatenation. We evaluate:

- 3-Channel input (no model modification):
  - $R_c||G_c||D_d$ (blue channel replaced)
  - $R_c||G_c||(B_c + D_d)/2$ (blue and depth averaged)
  - $(R_c + H_d)/2||(G_c + S_d)/2||(B_c + V_d)/2$ (RGB and HSV averaged)
- 4-Channel input (modified input layer):
  - $R_c||G_c||B_c||D_d$ (image and grayscale depth map concatenated)
  - The new 4th-channel weights are initialized as the average of the original RGB weights.
- 6-Channel input (modified input layer):
  - $R_c||G_c||B_c||H_d||S_d||V_d$ (image and colored depth map concatenated)
  - The weights for the three new channels (4-6) are initialized as a copy of the original RGB channel weights (1-3).

3.4.2. Late fusion

In this approach, the RGB image and the preprocessed depth map (i.e., the 3-channel false-color map) are processed independently by two instances of the same frozen VPR model. This produces two separate descriptors: d_{image} and d_{depth}. These descriptors are then merged to create the final descriptor, d_{fused}. We evaluate the following late fusion techniques:

- Non-learnable methods:
  - Concatenation: $d_{fused} = [d_{image}||d_{depth}]$. The final descriptor size is $512 + 512 = 1024$.
  - Element-wise sum: $d_{fused} = d_{image} + d_{depth}$. The size of the descriptor remains 512.
  - Weighted sum: $d_{fused} = w \cdot d_{image} + (1 - w) \cdot d_{depth}$. Various values for the weight w are tested.
  - Principal Component Analysis (PCA): The d_{image} and d_{depth} descriptors are first concatenated (1024) and a PCA model is fitted to the training data. The descriptors are then projected onto a lower-dimensional space (e.g., 256).
- Learnable methods:
  - Single-Layer Perceptron (SLP): A single fully-connected layer. This learns a more complex weighted sum of the concatenated descriptor with size 1024, outputting a descriptor with size d_{fused}.
  - Multi-Layer Perceptron (MLP): A small network of fully-connected layers, allowing for a more complex, non-linear transformation of the concatenated features into the final d_{fused}.

With regard to the monolayer and the multilayer perceptrons, they have been trained with 1000 training examples (Section 4.2.1 describes how the training samples are built and the rest of the implementation details). They receive the image and the depth descriptors concatenated as inputs and they output a new descriptor that combines both visual and depth information.

Table 1
Image sets employed for training and evaluation from the COLD database.

Environment	Train/Database Cloudy	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

* Training set.

4. Experiments

In this section, a comprehensive set of experiments is conducted to validate our central hypothesis: fusing visual data with geometric features obtained from depth estimators can significantly enhance VPR robustness. For this purpose, the dataset is first presented in Section 4.1. Second, the implementation details are listed in Section 4.2. Finally, the experimental results are divided into the ablation study (Section 4.3) and further evaluation (Section 4.4).

4.1. Dataset

The COLD database [52] is employed for our primary evaluation.[1] It was captured in indoor facilities of two universities: the University of Freiburg (FR) and the University of Saarbrücken (SA). Different sequences are captured at each building, which cover different illumination conditions: cloudy, night and sunny.

As detailed in Table 1, the database (map) sequences consist of cloudy images. The test (query) sequences include cloudy (same condition), night and sunny (cross-condition). This setup allows us to rigorously evaluate robustness to appearance changes. As for the learnable methods, a small training set is built exclusively from the FR-A cloudy sequence.

4.2. Implementation details

4.2.1. Training protocol

For the experiments that involve training, we employ a supervised contrastive triplet learning approach. We use an SGD (Stochastic Gradient Descent) optimizer (learning rate = 0.001) and the Batch Hard Loss [53], with a margin $m = 0.25$ and batch size $N = 16$. Positive/negative samples are determined by a distance threshold of $r = 0.4m$. All experiments were conducted on an NVIDIA GeForce RTX 4080 SUPER GPU.

4.2.2. Evaluation metrics

Performance is measured using the standard VPR metrics Recall@1 ($R@1$) and Recall@1% ($R@1\%$).

[1] More information can be found on their official project website https://www.cas.kth.se/COLD/.

- $R@1$: The percentage of queries for which the single best-matched database image is within a geometric error threshold d.
- $R@1\%$: The percentage of queries for which at least one of the top 1% retrieved database candidates is within the threshold d.

Following the COLD benchmark, a strict localization threshold of $d = 0.5$ m is used.

4.3. Ablation study

This section is structured into three different experiments: depth preprocessing (Section 4.3.1), early fusion (Section 4.3.2) and late fusion evaluation (Section 4.3.3). For the ablation study, the CosPlace model [18] is employed as the default VPR encoder and DAv2 [10] is used as the MDE model. Unless otherwise specified, the weights of the VPR model are kept frozen to test the zero-shot capabilities of the method.

4.3.1. Depth preprocessing

In this experiment, the preprocessing techniques described in Section 3.2 are evaluated. To this end, the CosPlace model computes the embeddings from the depth maps and place recognition is performed without using the visual information.

First, the optimal way to represent the single-channel depth map as an input for the three-channel, VPR encoder, must be determined. To this end, VPR is performed using only the depth maps, processed in five different ways as described in Section 3.2. Table 2 includes the global $R@1$ and $R@1\%$ results on the COLD dataset.

The results in Table 2 reveal a clear trend. The information provided by the original depth maps is not correctly perceived by the model, which was originally trained with color images. However, applying a false-color (HSV) map (Fig. 3(f)) or, to a lesser extent, sharpening (Fig. 3(e)), significantly boosts performance. These techniques re-format the geometric data into a representation that the RGB-trained encoder can better understand, with the 3-channel false-color map yielding the best results. Based on this, we use the sharpened (1-channel) and HSV-colored (3-channel) maps for subsequent fusion experiments.

4.3.2. Early fusion

Next, we evaluated the early fusion methods detailed in Section 3.4.1. The results are presented in Table 3.

Without any fine tuning, all early fusion methods failed to outperform the baseline (RGB-only). This is expected, as the weights from the first layer of the model are optimized for natural RGB images, and concatenating new channels (like depth) disrupts this. While fine tuning (right side of Table 3) closes this gap and provides a certain improvement for some configurations (e.g., $R||G||D$), the overall performance gain is not pronounced.

Table 2
Evaluation of different inputs to the CosPlace model on the COLD database.

Model input	R@1 (%)	R@1% (%)
RGB	**84.00**	**94.85**
Raw depth maps	62.24	78.42
+Equalized histogram	67.52	83.07
+Inverted	67.69	83.84
+Sharpening	76.84	91.93
+False color (HSV)	78.71	92.33

Table 3
Performance of early fusion methods before and after training.

N. channels	Method	w\o training		w\ training	
		R@1 (%)	R@1% (%)	R@1 (%)	R@1% (%)
3	RGB	**84.00**	**94.85**	84.35	95.53
6	RGB||HSV	81.67	94.29	83.03	95.25
4	RGB||D	83.02	94.35	84.63	**95.66**
3	R||G||D	83.02	94.00	**84.71**	95.56
3	R||G||(B+D)/2	83.65	94.65	84.40	95.07
3	(RGB + HSV)/2	81.01	93.57	83.18	95.29

Table 4
Performance of late fusion methods before and after training.

Method	w\o training		w\ training	
	R@1 (%)	R@1% (%)	R@1 (%)	R@1% (%)
RGB	84.00	94.85	84.35	95.53
Concat	**85.78**	**95.92**	**86.51**	**96.97**
Sum	84.97	95.66	86.00	96.66
Weighted sum ($w = 0.6$)	85.12	95.64	85.90	96.54
PCA ($n = 256$)	85.75	95.90	86.47	96.95
MLP BottNeck	81.32	94.33	83.82	96.00
MLP InvBottNeck	83.61	95.22	86.03	96.80
MP 1024→1024	84.37	95.42	86.15	96.87

4.3.3. Late fusion

In this section, the different late fusion methods proposed in Section 3.4.2 are evaluated. Table 4 displays the $R@1$ and $R@1\%$ results achieved with each method before and after training. To perform training, two independent CosPlace models are fine tuned, one of the models with RGB images and the other one with depth maps, which are specific to each modality. It must be noted that the learnable methods, i.e., the perceptrons, are always trained. Therefore, the expression "w\o training" only refers to the VPR model.

In contrast with early fusion approaches, the late fusion methods (Table 4) show a significant and immediate improvement. The key finding is that simple, non-learnable fusion techniques (Concatenation and PCA) substantially outperform the non-fusion baseline (RGB), even with zero fine tuning. Concatenation proved to be the most effective, yielding the highest $R@1$ and $R@1\%$ (left side of Table 4). Interestingly, the learnable methods (e.g., MLPs) did not provide additional benefits, suggesting that the descriptors from the frozen models are already in a highly descriptive space. Given these results, we select concatenation without fine tuning as the default proposed fusion method owing to its high performance.

4.4. Further evaluation

Having established the optimal late fusion strategy, a deeper analysis is now conducted to validate its robustness and generalizability. From this point, we will refer to this optimal configuration as PDPR.

4.4.1. Robustness against illumination shifts

First, the invariance of PDPR against pronounced changes in appearance, mainly due to lighting variations, is analyzed. Fig. 4 depicts the performance of our fusion approach compared to the baseline (RGB without depth enhancement or fusion) across different lighting scenarios from the COLD dataset: cloudy, night and sunny. In this figure, both the $R@1$ and the $R@1\%$ are expressed as a percentage. Prior to analyzing the results, it must be noted that the database sequence from each environment was captured under cloudy conditions. Therefore, night and sunny tests are the most challenging conditions. Besides, for the FR-B and SA-A environments, the results are not provided for certain lighting conditions because there are no available sequences on the dataset website.

From Fig. 4, it can be noticed that PDPR outperforms the baseline in every scenario in terms of $R@1$ and $R@1\%$. Importantly, the performance gap is most pronounced under the more challenging sunny

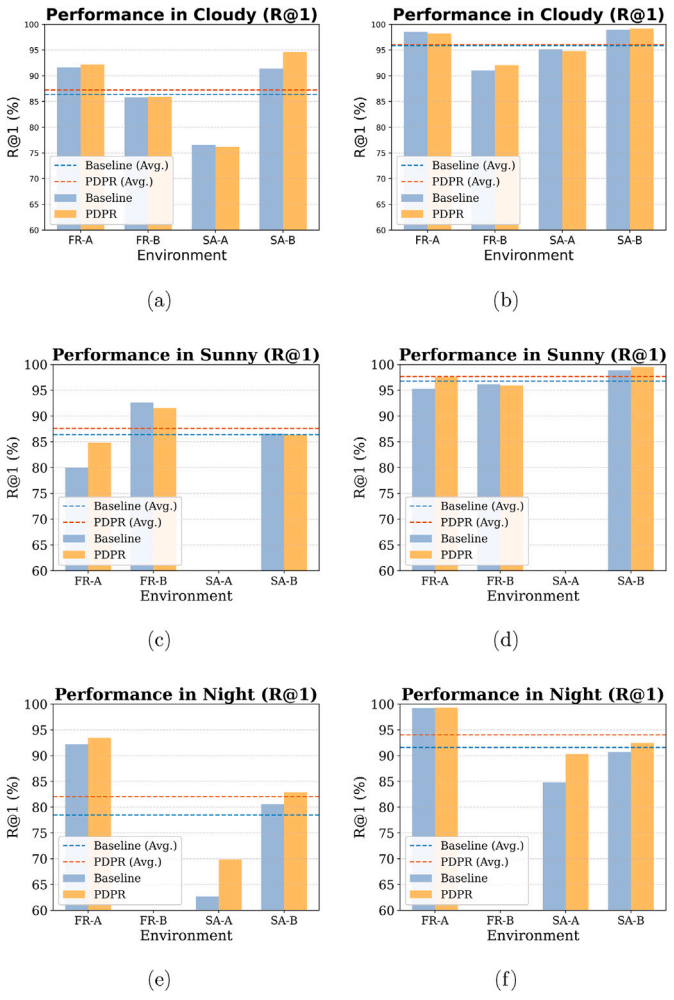


Fig. 4. Performance of PDPR in terms of $R@1$ and $R@1\%$ across diverse lighting conditions from the COLD dataset: cloudy (top), sunny (middle) and night (bottom).

(+1.18% $R@1$ without training and +1.76% with training) and night conditions (+3.58% $R@1$ without training and +4.33% with training), where standard VPR approaches struggle due to the cross condition.

It is important to note that the proposed method is not fully invariant to illumination changes, which is an inherent limitation of RGB-based VPR models due to their sensitivity to appearance shifts. However, the experimental results demonstrate that PDPR enhances the robustness of these approaches against lighting variations.

4.4.2. Robustness against noise

In this subsection, the robustness of PDPR against Gaussian noise is evaluated. The primary objective is to assess, qualitatively and quantitatively, whether the noise added to the RGB images is propagated to depth maps, and potentially compromises the performance of PDPR. To illustrate this, Fig. 5 includes several examples of a panoramic image subjected to varying levels of additive Gaussian noise (σ), whereas Fig. 6 shows the overall performance of PDPR with respect to the baseline CosPlace model (only RGB information) on the COLD dataset for every value of σ.

From Fig. 5, it can be noticed that the quality of depth maps remains fairly stable when Gaussian noise is added to the original image. A loss of detail becomes discernible only in the images when noise is pronounced ($\sigma \geq 20$). In quantitative terms (Fig. 6), a progressive decline in performance is observed. The method is barely affected by lower values of σ, and only drops significantly with very high values ($\sigma = 50$). It must be noted that such high noise values are very uncommon in real setups, so it is unlikely that noise becomes a major obstacle for the proposed approach. Crucially, the performance advantage over the baseline is maintained across the noise spectrum. This indicates that the proposed approach does not compromise the stability of the model, as it degrades in parallel with the original VPR model rather than becoming more sensitive.

4.4.3. Model-agnostic property

Next, the hypothesis that PDPR is model-agnostic is validated. The CosPlace encoder is replaced with a variety of different state-of-the-art VPR models (EigenPlaces, MixVPR, SALAD) and our zero-shot late fusion strategy is applied. Table 5 contains the results obtained with the baseline (RGB images are fed to the frozen VPR models) and with PDPR (late fusion between RGB images and depth maps).

Table 5 shows that PDPR consistently improves the performance of every model tested. This finding is critical, as it proves that the proposed framework is a solution that does not need to be intensely tuned for one specific architecture. Instead, it is a general, pluggable module that can be added to any pre-trained VPR model to enhance its robustness.

4.4.4. Evaluation with other datasets

To demonstrate generalizability beyond the COLD dataset, PDPR is evaluated with the 360Loc dataset [54].[2] This dataset contains equirectangular images from a 360 ° camera in a mixed indoor-outdoor campus environment, with challenging day/night sequences. Fig. 7 includes some examples of images from the 360Loc dataset and their corresponding depth maps obtained with DAv2 [10] and our preprocessing method.

To perform evaluation in this dataset, the *mapping* sequences are employed as the database (always captured under day conditions) and the *query_360* sequences are used as queries (captured under both day and night conditions). The test is conducted across all available sequences and lighting conditions. Fig. 8 displays a comparative evaluation between PDPR and the baseline (RGB-only). For inference, the CosPlace model is employed.

The results in Fig. 8 mirror our findings with the COLD dataset (please refer to Section 4.4.1). The proposed fusion approach provides a clear performance boost in most environments, especially in the difficult cross-illumination tests (i.e., “day” database vs. “night” query). This confirms that PDPR is robust not only to indoor lighting shifts but also to extreme day/night changes in outdoor scenarios.

4.4.5. Computational cost

In order to assess the feasibility of integrating this method into a real-time system, time and memory requirements are analyzed. Table 6 includes the memory requirements and inference time for every component of PDPR (late fusion).

As revealed in Table 6, the memory needed to run PDPR is fairly small (around 2.3GB in practice, considering internal calculations of the models), which is affordable for the majority of standard onboard devices. With regard to the time requirements, the average duration of a training process, which consists of 50k triplet samples seen by the model, is 285 s on an NVIDIA GeForce RTX 4080 SUPER GPU with 24GB of VRAM. Meanwhile, the inference time depends on the image resolution, ranging from 96 ms with the images from 360Loc to 249 ms with COLD. Although this latency is higher than that of highly efficient methods like CosPlace, it is still reasonable for real-time VPR applications that usually

[2] Additional information can be found on their official project website https://huajianup.github.io/research/360Loc/.

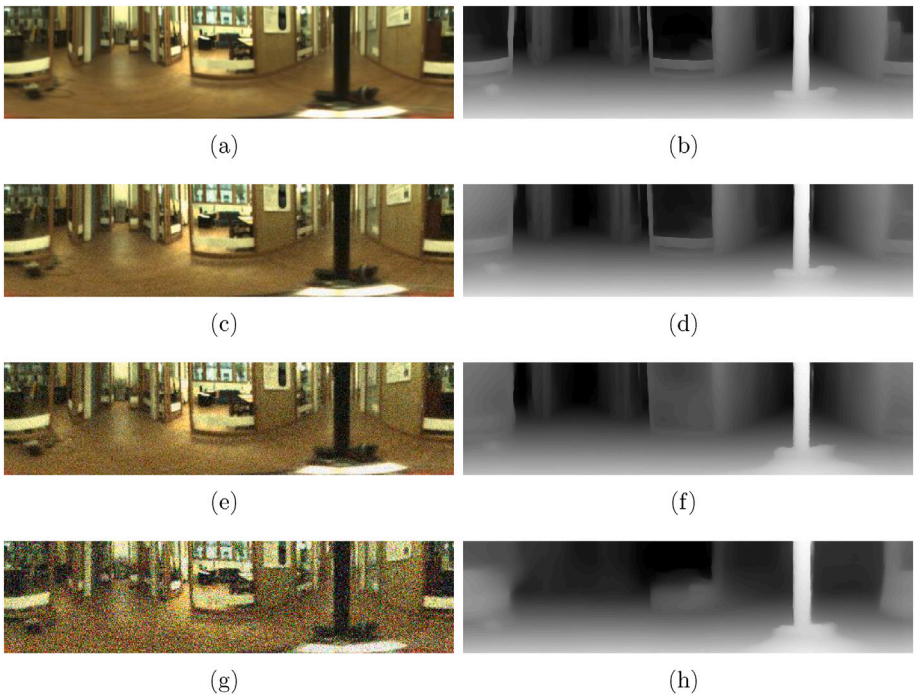


Fig. 5. Examples of an image from the COLD dataset (left) and its corresponding depth map (right) for different values (σ) of Gaussian noise: without noise (a,b), $\sigma = 10$ (c,d), $\sigma = 20$ (e,f) and $\sigma = 50$ (g,h).

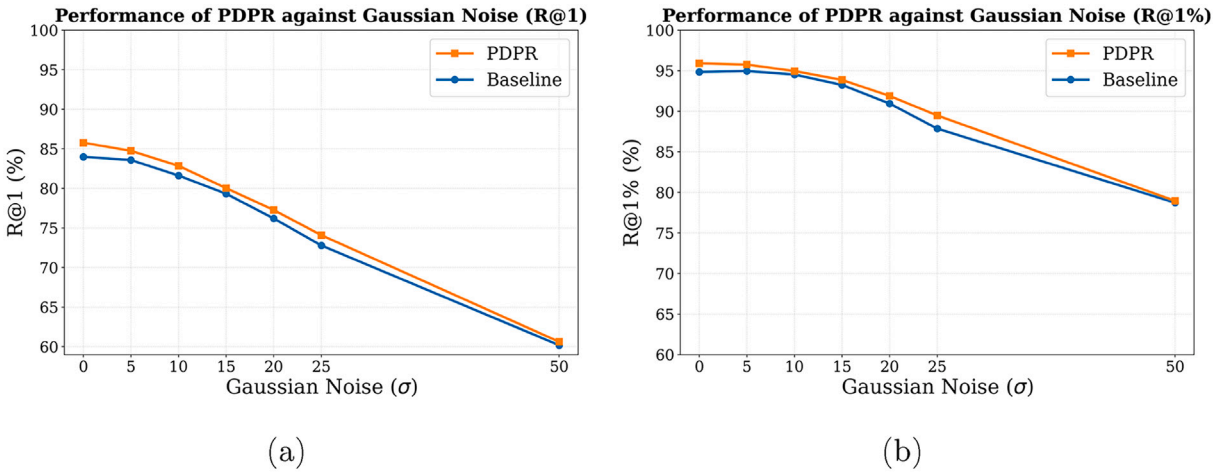


Fig. 6. Performance of PDPR in terms of $R@1$ (left) and $R@1\%$ (right) with Gaussian noise of different magnitude (σ).

work with a frequency of 2–10 Hz. The main bottleneck of the method is the depth estimator, but this issue can be tackled by replacing the model with one of its distilled versions (small or base). Fig. 9 shows the trade-off between inference speed and $R@1$ performance for these variants.

Fig. 9 reveals that the performance remains stable when using smaller versions of DAv2. Without fine tuning, there is a small decrease of 1.6% in $R@1$ on the COLD dataset when using the small and base models. At 360Loc, the base version exhibits a decrease of 2% relative to the large model, whereas the performance with DAv2-Small is around 3.4% lower in terms of $R@1$. In terms of inference speed, the base model and the small model are 2.3x and 4x faster than DAv2-Large, respectively. Consequently, for applications where high recall is required and inference speed is not critical, the large version of DAv2 is the most suitable choice. Otherwise, if a high-speed solution is needed, DAv2-Small and DAv2-Base achieve competitive accuracy with significantly lower latency.

Table 5
Evaluation of the performance of different VPR models without (left) and with PDPR (right).

Model	RGB		PDPR	
	R@1 (%)	R@1% (%)	R@1 (%)	R@1% (%)
CosPlace [18]	84.00	94.85	85.78 (+1.78%)	95.92 (+1.07%)
EigenPlaces [21]	84.55	94.97	85.74 (+1.19%)	96.10 (+1.13%)
MixVPR [22]	85.16	95.31	86.13 (+0.97%)	96.09 (+0.76%)
SALAD [23]	83.16	95.81	84.18 (+1.02%)	96.17 (+0.36%)

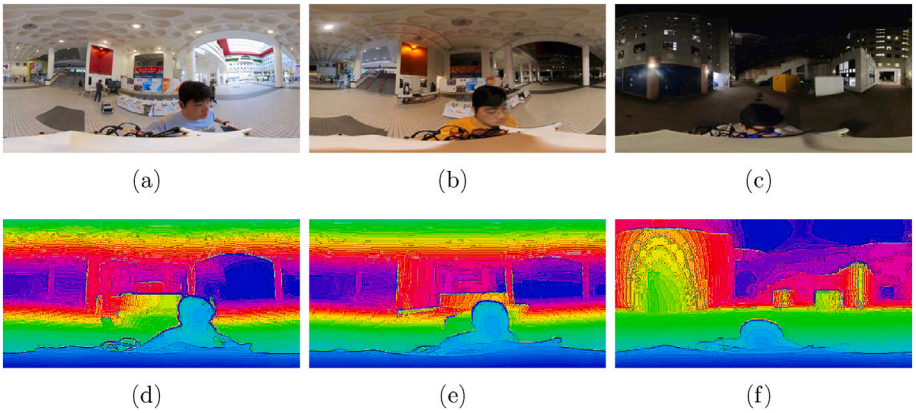


Fig. 7. Examples of (a,b,c) equirectangular images from the 360Loc dataset [54] and (d,e,f) their corresponding depth maps obtained with Depth Anything v2 [10] (on false color).

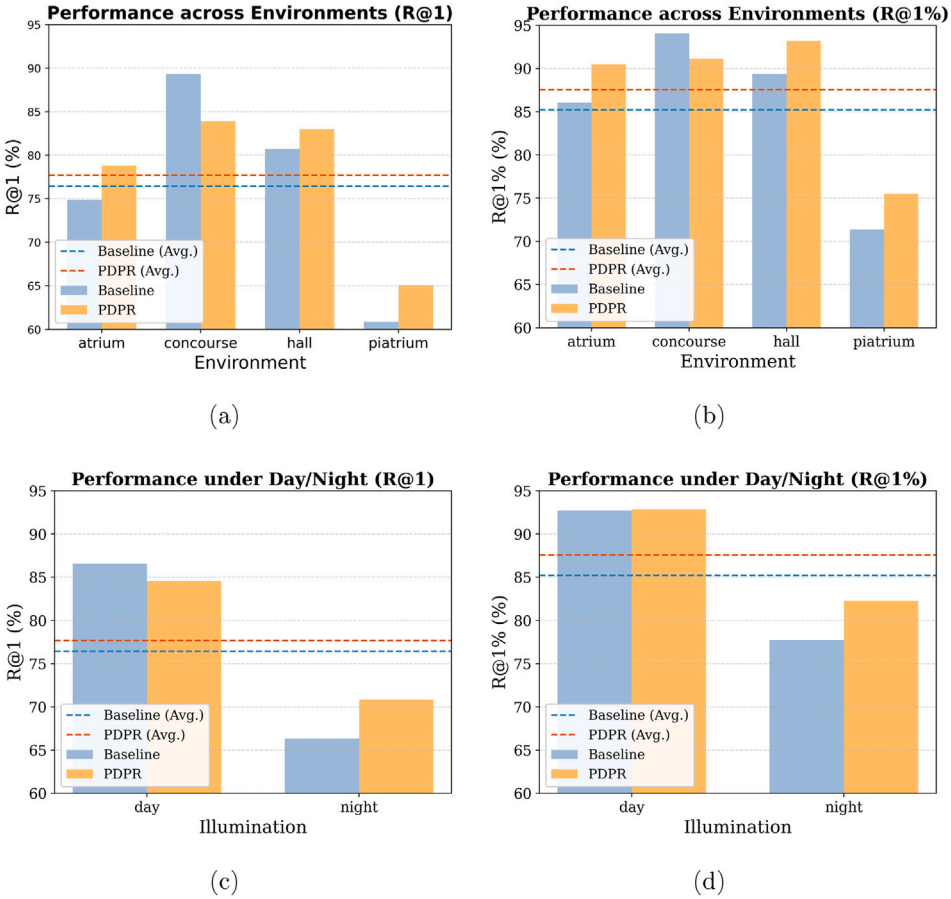


Fig. 8. Performance of PDPR in terms of $R@1$ (left) and $R@1\%$ (right) across diverse environments (top) and lighting conditions (bottom) in the 360Loc dataset.

Table 6
Computational cost of PDPR: inference time and memory requirements.

	Inference time		Memory
	COLD	360Loc	
DAv2-L	245 ms	92 ms	1.3GB
CosPlace	1 ms	1 ms	59MB
Image preprocessing	1 ms	1 ms	X
Query embedding 1x1024 (float32)	X	X	4KB
Database embeddings 1x1024 (float32)	X	X	2MB
Total PDPR time	249 ms	96 ms	1.36GB

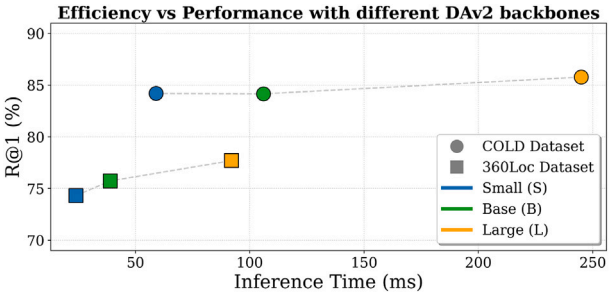


Fig. 9. Trade-off between inference speed and performance of PDPR with different DAv2 backbones (Small, Base and Large).

Table 7
Evaluation of different geometric-based inputs within the PDPR method on the COLD dataset.

Model input	R@1 (%)	R@1% (%)
Baseline (RGB)	84.00	94.85
Depth maps	**85.78**	**95.92**
Gradient magnitude	84.38	95.11
Surface normals	85.06	95.79

4.4.6. Analysis of depth features

The aim of this section is to establish and measure the benefits of adding depth information to the model. Additionally, we analyze the features that the models are extracting from the depth maps. First, a comparison among three different geometric-based inputs is performed: depth maps obtained with DAv2, gradient magnitude of the image and surface normals representation. Table 7 includes the $R@1$ and $R@1\%$ results of the late fusion between RGB images and the different geometric-based inputs.

Furthermore, we analyze the feature maps obtained by both CNN-based (e.g., CosPlace [18]) and transformer-based (e.g., SALAD [23]) VPR models using RGB images and depth maps. For the CosPlace model, EigenCAM is employed [55]. This tool projects the first PCA component of the feature maps from a specified layer onto the input image. The result is a heatmap that highlights the discriminative regions on which the model focuses. In our case, we apply EigenCAM to the last convolutional layer of the CosPlace models that have been trained with RGB images and depth maps (see Fig. 10). Concerning the SALAD model, since EigenCAM is not supported for transformer-based architectures, PCA with $n = 3$ is applied to the features from the last transformer block of DINOv2 [17], the SALAD encoder, and these PCA representations are depicted as false color maps. In these maps, regions with similar colors are semantically aligned for the model (see Fig. 11).

In the examples shown in Fig. 10, it can be observed that the model trained with RGB images focuses more on the textures of the ground, occasionally treating the robot chassis as an artifact. Conversely, the model that processes depth maps mainly focuses its attention on horizontal and vertical edges of doors or windows, and distinct elements like chairs or people. In Fig. 11, we notice that both modalities show a similar pattern distribution. The first component (red) corresponds mainly to the floor, the second (green) to the edges and the third (blue) to the objects and the background. However, several differences can be appreciated. First, a clearer scene layout can be observed from depth map PCA, as the color-based PCA aligns some textural information with the second component (edges). Second, in the depth-based PCA, the scene objects are more distinguishable due to their light blue color. In general terms, these PCA-based visualizations reinforce the initial hypothesis of this manuscript. The combination of visual and depth information for robust VPR is highly effective, as RGB images provide textural information from the scene, whereas depth maps contribute with geometric structure from specific objects and the general layout.

4.4.7. Qualitative results

To provide a clearer understanding of our method's robustness, we present several qualitative examples in Figs. 12 (COLD dataset) and Fig. 13 (360Loc dataset) where the baseline VPR model (CosPlace without fusion) fails, but PDPR successfully localizes the query. In these figures, a green frame indicates a correctly retrieved image, and red frames indicate an incorrect retrieval.[3]

From these examples, it can be observed that PDPR exhibits robustness against especially challenging situations, such as natural spotlights (Fig. 12, 2nd row) and shadows (Fig. 13, 2nd row), mirror reflections (Fig. 12, 3rd row), visual aliasing (Fig. 13, fourth row) or even extreme appearance shifts due to illumination (Fig. 13, 3rd row). These examples demonstrate the strength of our method against all these situations that frequently occur during the navigation of the vehicle.

4.4.8. Limitations

Finally, the limitations of the proposed method are discussed. Figs. 14 and 15 contain several examples from both the COLD and the 360Loc datasets, respectively, in which PDPR failed to retrieve the correct location.

From Figs. 14 and 15, it can be observed that PDPR struggles when faced with particularly challenging situations. In the first example of Fig. 14, an extreme ambiguity between the query and the retrieved position can be noticed, with the presence of highly similar objects in both images, and a relatively high difference between the query and the ground truth due to the different lighting conditions. In the second example of Fig. 14, poor conditions cause the query image to contain some noise, which is propagated into the generated depth map. In the first case of Fig. 15, dynamic elements are present. Large, unmapped dynamic elements (e.g., people, moved furniture) break the assumption that geometric information is always robust against appearance changes, introducing geometric noise that does not match the database. Finally, in the second example of Fig. 15, some artifacts are generated by the DAv2 model. Artificial illumination can cause the model to hallucinate and generate false structures (e.g., a "false ceiling"), introducing geometric errors that lead to a mismatch.

We note that most of these failures also occur in the baseline method, suggesting PDPR is not able to resolve the errors under extremely challenging conditions. However, this analysis confirms that the performance of our method is intrinsically linked to the performance of the original VPR model and the robustness of the depth estimator. This fact leads to the conclusion that the proposed method is not fully invariant to illumination shifts, as the proposed method still depends on RGB images and the generated depth maps are conditioned by the quality of these images. Nevertheless, the experiments demonstrate that the proposed approach improves the robustness of current VPR approaches. Besides,

[3] More qualitative results can be found on our project website https://marcosalfaro.github.io/projects-PDPR/.

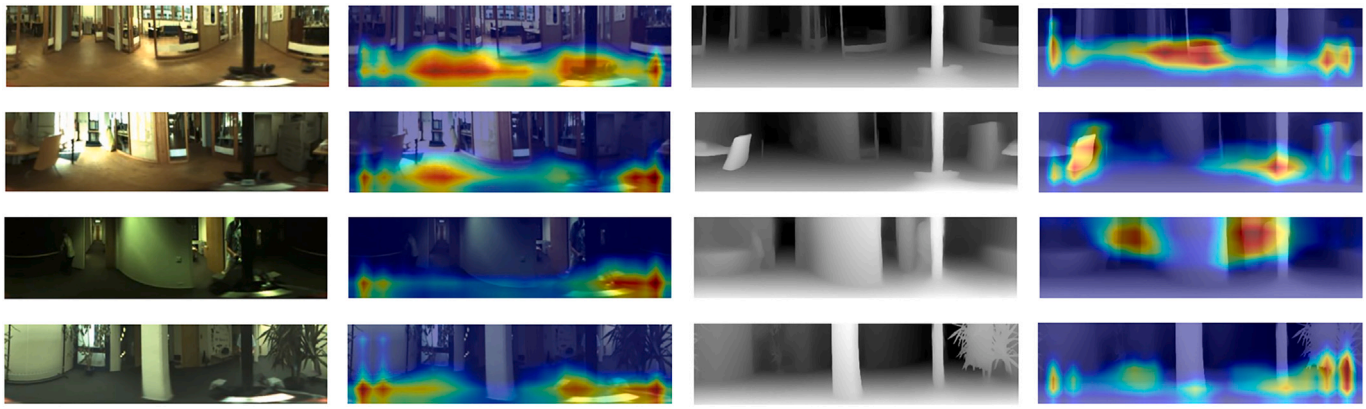


Fig. 10. EigenCAM representations of the last convolutional layer from CosPlace, obtained from RGB images (left) and depth maps (right).

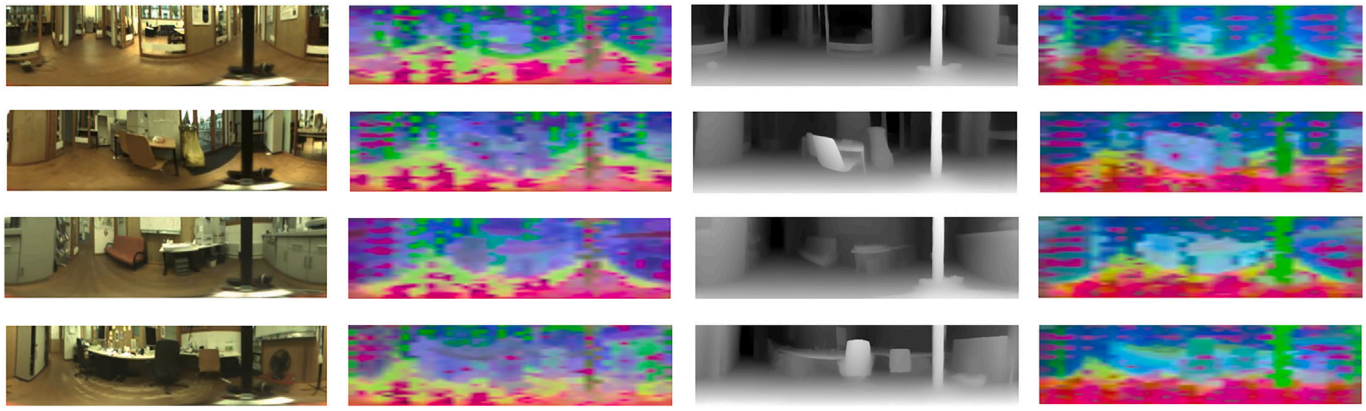


Fig. 11. PCA representations of the last transformer block from SALAD, obtained from RGB images (left) and depth maps (right).

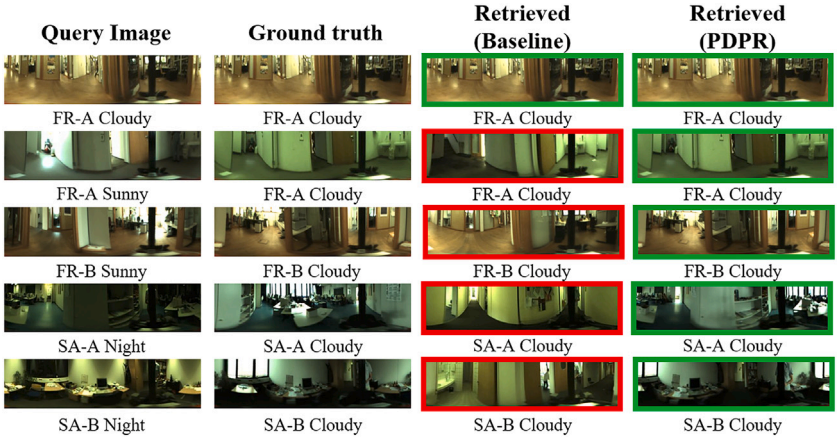


Fig. 12. Qualitative results from the COLD dataset.

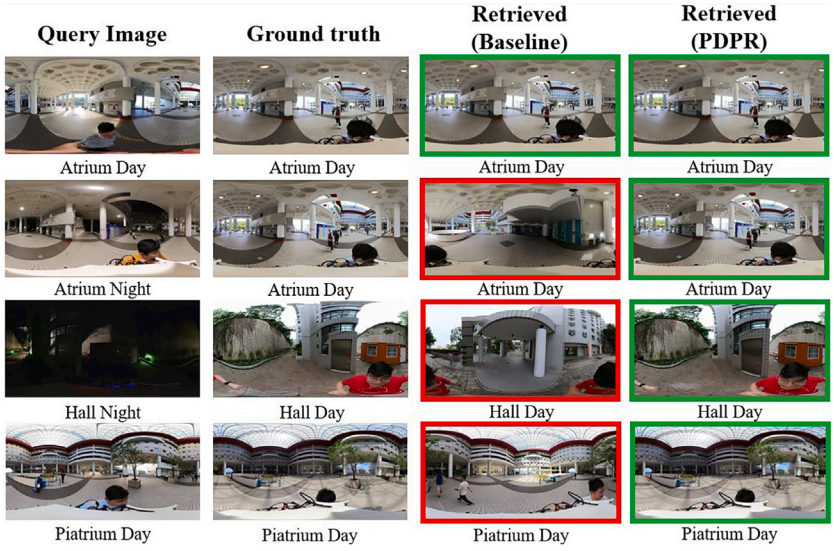


Fig. 13. Qualitative results from the 360Loc dataset.

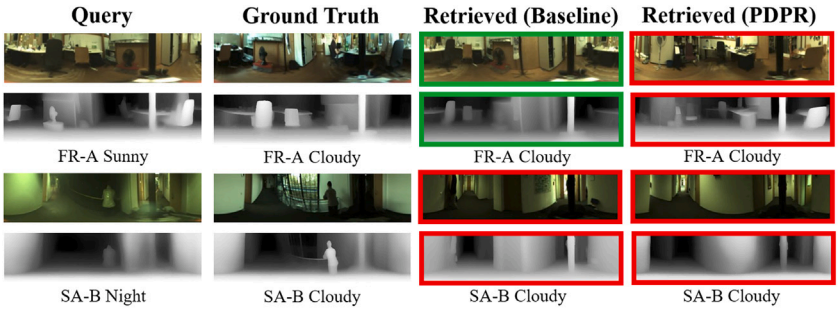


Fig. 14. Several examples from the COLD dataset where PDPR fails to retrieve the query image.

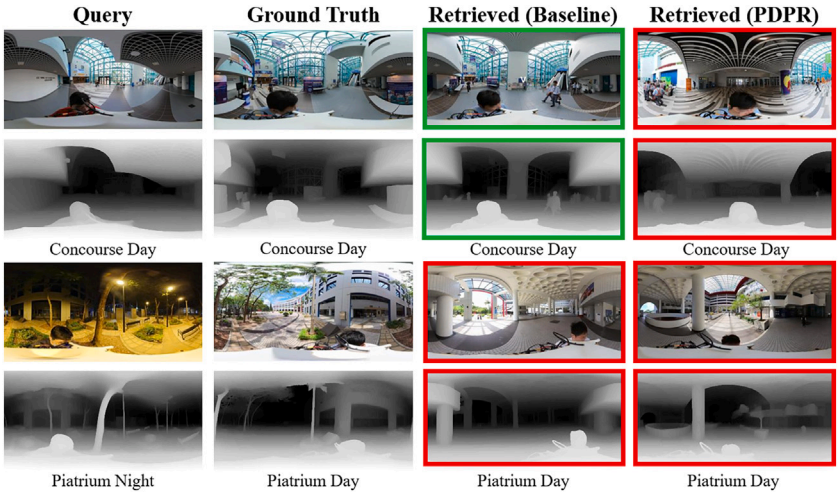


Fig. 15. Several examples from the 360Loc dataset where PDPR fails to retrieve the query image.

with future developments in the MDE field, the performance of PDPR will improve as well.

5. Conclusions

This paper demonstrates that fusing visual data with MDE-derived geometric features provides a significant, low-cost enhancement to VPR robustness. The experiments have shown that PDPR, acting as a pluggable, zero-shot module, consistently improves the retrieval performance of multiple state-of-the-art VPR models. Moreover, this approach is particularly effective at increasing their robustness against pronounced illumination changes. Our ablation studies also confirmed that proper depth-map preprocessing (e.g., false-color mapping) is a critical step for aligning geometric data with RGB-trained encoders. Besides, late fusion approaches have proven to be more effective than the early approaches.

The experimental results also highlight that the fusion performance is intrinsically bound to the quality of the MDE model. As demonstrated in our limitations analysis, MDE failures from noisy images or poor illumination conditions propagate into localization errors. This points to a clear and promising direction for future work: uncertainty-aware fusion. By integrating the uncertainty of the MDE model output, a future framework could learn to adaptively down-weight the depth descriptor when it is predicted to be unreliable (e.g., in poor-quality outputs or highly dynamic scenes), thus focusing more on the visual descriptor.

In conclusion, PDPR provides a practical path for robust robotic localization without the cost and complexity of additional hardware devices. The principles demonstrated here are not limited to VPR; future work will explore the integration of visual and MDE-derived geometric data for other navigation tasks, such as 3D object detection and traversable-area segmentation.

CRediT authorship contribution statement

Marcos Alfaro: Writing – original draft, Visualization, Software, Investigation, Data curation. **Juan José Cabrera:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis **Arturo Gil:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization. **Oscar Reinoso:** Visualization, Validation, Methodology, Formal analysis, Data curation. **Luis Payá:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The Ministry of Science, Innovation and Universities (Spain) has supported this work through “Ayudas para la Formación de Profesorado Universitario” (M. Alfaro, FPU23/00587; J.J. Cabrera, FPU21/04969). This research work is part of the project PID2023-149575OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE. It is also part of the project CIPROM/2024/8, funded by Generalitat Valenciana, Conselleria de Educación, Cultura, Universidades y Empleo (program PROMETEO 2025).

Data availability

The code used in the experiments is available at https://marcosalfaro.github.io/projects-PDPR/. The images from the COLD and 360Loc databases can be downloaded from: https://www.cas.kth.se/COLD/, https://github.com/HuajianUP/360Loc.

References

[1] J. Komorowski, Minkloc3D: point cloud based large-scale place recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1790–1799, https://doi.org/10.48550/arXiv.2011.04530
[2] S. Schubert, P. Neubert, S. Garg, M. Milford, T. Fischer, Visual place recognition: a tutorial. IEEE Robot. &Autom. Mag. 31 (3) (2023) 139–153, https://doi.org/10.1109/MRA.2023.3310859
[3] M. Flores, D. Valiente, A. Peidró, O. Reinoso, L. Payá, Generating a full spherical view by modeling the relation between two fisheye images, Vis. Comput. (2024) 1–26, https://doi.org/10.1007/s00371-024-03293-7
[4] O. Siméoni, H.V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, et al. arXiv preprint arXiv:2508.10104, 2025, https://doi.org/10.48550/arXiv.2508.10104
[5] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1655–1668, https://doi.org/10.1109/TPAMI.2018.2846566
[6] S. Hangloo, B. Arora, Multimodal fusion techniques: review, data representation, information fusion, and application areas, Neurocomputing (2025) 130827, https://doi.org/10.1016/j.neucom.2025.130827
[7] H. Lai, P. Yin, S. Scherer, Adafusion: Visual-LiDAR fusion with adaptive weights for place recognition.IEEE, Robot. Autom. Lett. 7 (4) (2022) 12038–12045, https://doi.org/10.1109/LRA.2022.3210880
[8] D. Yudin, Y. Solomentsev, R. Musaev, A. Staroverov, A.I. Panov, HPointLoc: point-based indoor place recognition using synthetic RGB-D images. in: International Conference on Neural Information Processing, Nov 2022, pp. 471–484, https://doi.org/10.1007/978-3-031-30111-7_40
[9] F. Endres, J. Hess, J. Sturm, D. Cremers, W. Burgard, 3-D mapping with an RGB-D camera, IEEE Trans. Robot. 30 (1) (2013) 177–187, https://doi.org/10.1109/TRO.2013.2279412
[10] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, H. Zhao, Depth anything v2.advances in neural, Inf. Process. Syst. 37 (2024) 21875–21911, https://doi.org/10.48550/arXiv.2406.09414
[11] L. Payá, F. Amorós, L. Fernández, O. Reinoso, Performance of global-appearance descriptors in map building and localization using omnidirectional vision, Sensors 14 (2) (2014) 3033–3064, https://doi.org/10.3390/s140203033
[12] D. Gálvez-López, J.D. Tardos, Bags of binary words for fast place recognition in image sequences, IEEE Trans. Robot. 28 (5) (2012) 1188–1197, https://doi.org/10.1109/TRO.2012.2197158
[13] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, NetVLAD: CNN architecture for weakly supervised place, in: recognition.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307, https://doi.org/10.48550/arXiv.1511.07247
[14] A. Dosovitskiy, An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020, https://doi.org/10.48550/arXiv.2010.11929
[15] N. Keetha, A. Mishra, J. Karhade, K.M. Jatavallabhula, S. Scherer, M. Krishna, S. Garg, AnyLoc: towards universal visual place recognition, IEEE Robot. Autom. Lett. (2023), https://doi.org/10.1109/LRA.2023.3343602
[16] C. Ryali, Y.T. Hu, D. Bolya, C. Wei, H. Fan, P.Y. Huang, C. Feichtenhofer, Hiera: a hierarchical vision transformer without the bells-and-whistles, in: International Conference on Machine Learning, Jul 2023, pp. 29441–29454, https://doi.org/10.48550/arXiv.2306.00989
[17] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Bojanowski). arXiv preprint arXiv:2304.07193, 2023, https://doi.org/10.48550/arXiv.2304.07193
[18] G. Berton, C. Masone, B. Caputo, Rethinking visual geo-localization for large-scale applications.in, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4878–4888, https://doi.org/10.48550/arXiv.2204.02287
[19] N. Sünderhauf, P. Neubert, P. Protzel, Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons, in: Proc. Of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA) (p. 2013), May 2013, https://arxiv.org/pdf/1906.12176.
[20] A. Ali-Bey, B. Chaib-Draa, P. Giguere, GSV-cities: toward appropriate supervised visual place recognition, Neurocomputing 513 (2022) 194–203, https://doi.org/10.1016/j.neucom.2022.09.127
[21] G. Berton, G. Trivigno, B. Caputo, C. Masone, Eigenplaces: training viewpoint robust models for visual place, in: recognition.In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11080–11090, https://doi.org/10.48550/arXiv.2308.10832
[22] A. Ali-Bey, B. Chaib-Draa, P. Giguere, MixVPR: feature mixing for visual place recognition.in, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2998–3007, https://doi.org/10.48550/arXiv.2303.02190
[23] S. Izquierdo, J. Civera, Optimal transport aggregation for visual place recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17658–17668, https://doi.org/10.48550/arXiv.2311.15937
[24] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, C. Yuan, Towards seamless adaptation of pre-trained models for visual place recognition, arXiv preprint arXiv:2402.14505, 2024, https://doi.org/10.48550/arXiv.2402.14505
[25] T.H. Wang, H.J. Huang, J.T. Lin, C.W. Hu, K.H. Zeng, M. Sun, Omnidirectional CNN for visual place recognition and navigation, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 2341–2348, https://doi.org/10.1109/ICRA.2018.8463173

[26] J.J. Cabrera, V. Román, A. Gil, O. Reinoso, L. Payá, An experimental evaluation of siamese neural networks for robot localization using omnidirectional imaging in indoor environments. Artif. Intell. Rev. 57 (8), Springer, (2024), p. 198, https://doi.org/10.1007/s10462-024-10840-0.
[27] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer, IEEE Trans. Pattern Anal. Mach. Intell. 44 (3) (2020) 1623–1637, https://doi.org/10.1109/TPAMI.2020.3019967
[28] J. Mao, D. Yao, Y. Hu, S. Chan, W. Sheng, H. Qin, DSSA-depth: unsupervised monocular depth estimation method based on Dual-Scale Self-Attention, Neurocomputing (2025) 132089, https://doi.org/10.1016/j.neucom.2025.132089
[29] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, et al., Metric3D v2: a versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation, IEEE Trans. Pattern Anal. Mach. Intell. (2024), https://doi.org/10.1109/TPAMI.2024.3444912
[30] B. Ke, A. Obukhov, S. Huang, N. Metzger, R.C. Daudt, K. Schindler, Repurposing diffusion-based image generators for monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9492–9502, https://doi.org/10.48550/arXiv.2312.02145
[31] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, X. Long, Geowizard: unleashing the diffusion priors for 3D geometry estimation from a single image, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, Sep 2024, pp. 241–258, https://doi.org/10.1007/978-3-031-72670-5_14
[32] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, H. Zhao, Depth anything: unleashing the power of large-scale unlabeled data, in: InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10371–10381, https://doi.org/10.48550/arXiv.2401.10891
[33] Z. Cao, J. Zhu, W. Zhang, H. Ai, H. Bai, H. Zhao, L. Wang, PanDA: towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 982–992, https://doi.org/10.48550/arXiv.2406.13378
[34] H. Li, W. Zheng, J. He, Y. Liu, X. Lin, X. Yang, et al., DA ^{2}: Depth Anything in Any Direction, arXiv preprint arXiv:2509.26618, 2025, https://doi.org/10.48550/arXiv.2509.26618
[35] S. Zhao, M. Gong, H. Fu, D. Tao, Adaptive context-aware multi-modal network for depth completion.IEEE, Trans. Image Process. 30 (2021) 5264–5276, https://doi.org/10.1109/TIP.2021.3079821
[36] N. Keetha, N. Müller, J. Schönberger, L. Porzi, Y. Zhang, T. Fischer, et al., MapAnything: Universal Feed-Forward Metric 3D Reconstruction, arXiv preprint arXiv:2509.13414, 2025, https://doi.org/10.48550/arXiv.2509.13414
[37] M. Kiray, A. Karaomer, B. Busam, Dropping the D: RGB-D SLAM Without the Depth Sensor, arXiv preprint arXiv:2510.06216, 2025, https://doi.org/10.48550/arXiv.2510.06216
[38] Y. Lu, Q. Wang, S. Ma, T. Geng, Y.V. Chen, H. Chen, D. Liu, Transflow: transformer as flow learner, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18063–18073, https://doi.org/10.48550/arXiv.2304.11523
[39] C. Han, Y. Lu, G. Sun, J.C. Liang, Z. Cao, Q. Wang, et al.). arXiv preprint arXiv:2406.01559, 2024, https://doi.org/10.48550/arXiv.2406.01559
[40] Z. Li, T. Shang, P. Xu, Z. Deng, Place recognition meet multiple modalities: a comprehensive review, current challenges and future development, Artif. Intell. Rev. 58 (11) (2025) 363, https://doi.org/10.1007/s10462-025-11367-8
[41] S. Shubodh, M. Omama, H. Zaidi, U.S. Parihar, M. Krishna, Lip-Loc: LiDAR image pretraining for cross-modal localization.in, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 948–957, https://doi.org/10.48550/arXiv.2312.16648
[42] H. Zhu, J.B. Weibel, S. Lu, Discriminative multi-modal feature fusion for RGB-d indoor scene recognition.in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2969–2976, https://doi.org/10.1109/CVPR.2016.324
[43] W. Liu, J. Fei, Z. Zhu, MFF-PR: point cloud and image multi-modal feature fusion for place recognition.in 2022, in: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Oct 2022, pp. 647–655, https://doi.org/10.1109/ISMAR55827.2022.00082
[44] Z. Zhou, J. Xu, G. Xiong, J. Ma, LCPR: a Multi-Scale Attention-Based LiDAR-Camera fusion network for place recognition, IEEE Robot. Autom. Lett. (2023), https://doi.org/10.1109/LRA.2023.3346753
[45] Y. Pan, J. Xie, J. Wu, B. Zhou, Camera-LiDAR fusion with latent correlation for Cross-Scene place recognition, IEEE Trans. Ind. Electron. (2024), https://doi.org/10.1109/TIE.2024.3440470
[46] W. Zhou, J. Liu, J. Lei, L. Yu, J.N. Hwang, GMNet: graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation. IEEE Trans. Image Process. 30 (2021) 7790–7802, https://doi.org/10.1109/TIP.2021.3109518
[47] J. Komorowski, M. Wysoczańska, T. Trzcinski, MinkLoc++: LiDAR and monocular image fusion for place recognition.in 2021, in: International Joint Conference on Neural Networks (IJCNN), Jul 2021, pp. 1–8, https://doi.org/10.1109/IJCNN52387.2021.9533373
[48] D. Cattaneo, M. Vaghi, S. Fontana, A.L. Ballardini, D.G. Sorrenti, Global visual localization in LiDAR-maps through shared 2D-3D embedding space, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), May 2020, pp. 4365–4371, https://doi.org/10.1109/ICRA40945.2020.9196859
[49] Z. Zhao, H. Yu, C. Lyu, W. Yang, S. Scherer, Attention-enhanced cross-modal localization between spherical images and point clouds, IEEE Sens. J. (2023), https://doi.org/10.1109/JSEN.2023.3306377
[50] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12179–12188, https://doi.org/10.48550/arXiv.2103.13413
[51] K. Simonyan, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014, https://doi.org/10.48550/arXiv.1409.1556
[52] A. Pronobis, B. Caputo, COLD: the CoSy localization database, Int. J. Robot. Res. 28 (5) (2009) 588–594, SageJournals, https://doi.org/10.1177/0278364909103912
[53] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737, 2017, https://doi.org/10.48550/arXiv.1703.07737
[54] H. Huang, C. Liu, Y. Zhu, H. Cheng, T. Braud, S.K. Yeung, 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, 2024, pp. 22314–22324, https://doi.org/10.48550/arXiv.2311.17389
[55] M.B. Muhammad, M. Yeasin, Eigen-cam: class activation map using principal components, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, Jul 2020, pp. 1–7, https://doi.org/10.1109/IJCNN48605.2020.9206626

Author biography

Marcos Alfaro is an Electronic and Industrial Automation Engineer and holds a master's degree in Robotics. He is a PhD candidate at the Miguel Hernández University with the national scholarship "Training of University Teachers" (2024-now). He is a member of the "Automation, Robotics and Computer Vision Lab" (2023-now). His research is focused on deep learning, computer vision and mobile robotics. In particular, his current research consists in developing algorithms for visual and multi-modal place recognition. He has been a visiting researcher at the Visual and Multimodal Applied Learning Lab at the Politecnico di Torino (2025).

Dr. Juan José Cabrera is an Electronic and Industrial Automation Engineer and holds a master's degree in Robotics. He is a post-doctoral student at the Miguel Hernández University with the national scholarship "Training of University Teachers" (2025-now). He is a member of the "Automation, Robotics and Computer Vision Lab" (2020-now). His research is focused on deep learning, computer vision and mobile robotics. In particular, he has developed algorithms for visual, LiDAR, multimodal and aerial place recognition. He has been a visiting researcher at the University of Coimbra (2023), University of Oxford (2024) and Polytechnic of Montreal (2025).

Prof. Arturo Gil is an Industrial Engineer and holds a PhD in Industrial Technologies. Since 2024, he is a Full Professor at the Miguel Hernández University. He conducts his research at the "Automation, Robotics and Computer Vision Lab". His main topics include deep learning, computer vision, data fusion and robotics. He has led several national and regional projects, which include the development of deep learning models for localization, mapping and scene understanding in surveillance and security applications. He has been a visiting researcher at the University of Technology Sydney and the Albert-Ludwigs University of Freiburg.

Prof. Oscar Reinoso is an Industrial Engineer and holds a PhD in Industrial Technologies. Since 2011, he is a Full Professor at the Miguel Hernández University. He conducts his research at the "Automation, Robotics and Computer Vision Lab". His main topics include deep learning, computer vision, intelligent systems and robotics. He has led several national and regional projects, which include the design of navigation, recognition and manipulation algorithms for the integration of intelligent robots into society, or the design of hybrid robots and multi-sensory reconstruction for applications in reticular structures.

Prof. Luis Payá is an Industrial Engineer and holds a PhD in Industrial Technologies. Since 2023, he is a Full Professor at the Miguel Hernández University. He conducts his research at the "Automation, Robotics and Computer Vision Lab". His main topics include deep learning, computer vision, data fusion and robotics. He has led several national and regional projects in the field of mobile robotics for surveillance and security tasks or hybrid robots for applications in reticular structures, among others. He has been a visiting researcher at the University of Bristol and the Imperial College London.