Expert Systems With Applications 328 (2026) 132838

Contents lists available at ScienceDirect

Expert Systems With Applications

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

CrossPlace: Cross-modal place recognition between fisheye cameras and LiDAR via a unified descriptor space

Juan José Cabrera [a,*], Marcos Alfaro [a], María Flores [a], Álvaro Martínez [a], Arturo Gil [a], Luis Payá [a,b]

[a] Institute for Engineering Research (I3E), Miguel Hernández University, Elche, Alicante, Spain
[b] Valencian Graduate School and Research Network for Artificial Intelligence (valgrAI), Valencia, Spain

ARTICLE INFO

Keywords:
Place recognition
LiDAR
Fisheye cameras
Intensity
Depth
Semantic segmentation

ABSTRACT

This paper presents CrossPlace, an innovative method for cross-modal place recognition between heterogeneous sensor modalities, particularly between fisheye cameras and LiDAR. Place recognition is the fundamental capability of mobile robots to determine their most likely location within a database, based on sensory input queries. In cross-modal place recognition, the goal is to localize using a different sensor from the one originally used to construct the database. The core contribution of this paper is a unified feature space that integrates intensity, depth and semantic information. Both the database entries and the queries are obtained by embedding sensor readings through the same CrossPlace model, ensuring a consistent representation across modalities. Consequently, a database constructed from LiDAR can be queried with fisheye images, and vice versa, using a single shared architecture. Furthermore, a comprehensive data transformation and preprocessing pipeline is presented. Specifically, CrossPlace is constituted by three independent branches, each one for processing intensity, depth and semantic information. Each branch consists of a CosPlace model for image embedding with shared weights across sensor modalities. Late fusion through concatenation of the intensity, depth and semantic embeddings provides optimal global performance. We conduct an exhaustive evaluation on the KITTI-360 dataset, where CrossPlace surpasses state-of-the-art techniques across all metrics, establishing a new standard for cross-modal place recognition in urban and highway environments. The results demonstrate the effectiveness of our unified approach for place recognition across different sensor modalities while maintaining a robust performance under various operating environments.

1. Introduction

Traditionally, place recognition has been carried out using a unique sensor modality, whether camera, LiDAR or Radar, both to capture the database and to acquire query observations during navigation. This approach is effective and has proven to be robust across diverse environmental and structural conditions (Wang et al., 2024b; Zhou et al., 2025). On some occasions, the sensor used to capture the dataset and the sensor mounted on the mobile platform (i.e. the one that captures the queries) have different specifications. This may happen, for example, when the initial sensor is substituted by a sensor with a higher resolution, greater precision, or a longer range. This situation is generally manageable because the data can be filtered to match the original database (Guan et al., 2023; Jung et al., 2025). This paper focuses on a more general and challenging scenario: the database is built with one sensing modality, while localization must be carried out using a completely different sensor type. This technique avoids the high costs of capturing and storing new datasets with the new sensor if the initial sensor on the mobile platform is replaced. Additionally, multi-robot systems may have different sensor configurations to operate in the same environment, making it necessary to develop methods that enable place recognition across different sensor modalities, such as cameras and LiDARs.

The approach presented in this study has great practical potential, as it eliminates the need to maintain the same type of sensor for both database and query readings. This enables the incorporation of more advanced sensors like LiDARs or new robots to the system with different sensor configurations, eliminating the burden of capturing new complete datasets. It also opens the possibility of creating a database using a LiDAR sensor and then using less expensive RGB cameras for day-to-day operation.

* Corresponding author.
E-mail addresses: juan.cabreram@umh.es (J.J. Cabrera), malfaro@umh.es (M. Alfaro), m.flores@umh.es (M. Flores), alvaro.martinezb@umh.es (Martínez), arturo.gil@umh.es (A. Gil), lpaya@umh.es (L. Payá).

https://doi.org/10.1016/j.eswa.2026.132838
Received 24 November 2025; Received in revised form 17 April 2026; Accepted 9 May 2026
Available online 15 May 2026

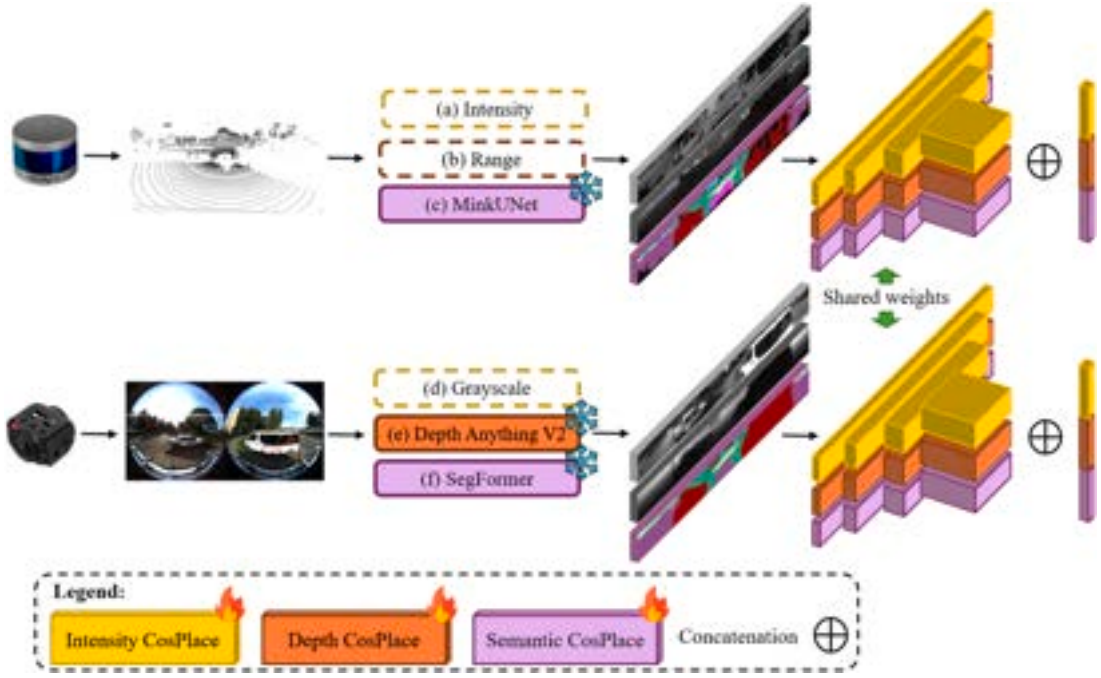

Fig. 1. General architecture of the CrossPlace method. LiDAR and fisheye inputs are transformed into intensity (a, d), depth (b, e) and semantic (c, f) images. Snowflake icons indicate frozen preprocessing models (MinkUNet34C Choy et al., 2019, Depth Anything V2 Yang et al., 2025, SegFormer Xie et al., 2021), while fire icons denote trainable CosPlace backbones (Berton et al., 2022). These backbones operate both with camera and LiDAR by sharing weights between sensor modalities to enforce a unified latent space. The final embedding is the concatenation of all branches.

This research proposes a method for place recognition between LiDAR sensors and omnidirectional vision systems (composed of two fisheye cameras, which are rotated 180° from each other). To achieve this, readings from different sensor types must be transformed into the same descriptor space. Specifically, both omnidirectional views (from the fisheye cameras) and LiDAR scans are transformed into intensity, depth and semantic spaces, enabling the use of the same network architecture for both types of sensor data. In summary, the main contributions of this approach are:

- **CrossPlace:** a novel method for place recognition between heterogeneous sensor modalities (LiDAR and omnidirectional fisheye cameras), by transforming both sensor inputs into a shared feature space of intensity, depth and semantics. Thus, CrossPlace enables the use of the same neural network to perform place recognition regardless of the input sensor data.
- **Modality transformation and representation optimization:** a comprehensive processing pipeline that converts fisheye images and LiDAR point clouds into unified intensity, depth and semantic representations. This includes advanced techniques such as depth estimation, semantic segmentation and LiDAR data completion.
- **Unified training and feature fusion strategy:** a unified cross-modal training approach is proposed, which randomly selects similar and different pairs of scenes captured with the same or different sensor types. As a result, it enables the model to learn a common descriptor space while avoiding the use of distillation techniques.
- **Comprehensive evaluation:** a detailed analysis of the performance of CrossPlace on the KITTI-360 dataset is provided, demonstrating the feasibility of place recognition with different sensors in urban and highway environments. The proposed approach significantly outperforms existing methods, setting a new standard for cross-modal place recognition between LiDAR and omnidirectional vision systems (based on two fisheye cameras rotated 180° from each other).

2. Related work

Cross-modal place recognition is an emerging area in mobile robotics. Traditionally, place recognition systems have used the same type of sensor to capture both the database and the query readings used to carry out place recognition. However, this usual approach may present practical limitations as sensors can change over time. In addition, it could be beneficial to adapt recognition systems to new sensor modalities while avoiding the need to capture a new database. In this context, it is necessary to develop methods that enable place recognition using different sensor modalities, such as cameras and LiDARs.

In recent years, various approaches have emerged to address cross-modal place recognition. Cattaneo et al. (2020) focus on bringing standard images and point clouds to the same descriptor space using VGG16 (Simonyan & Zisserman, 2014) and PointNet (Qi et al., 2017), respectively. Moreover, Yin et al. (2021) propose i3dLoc, a method that seeks robustness against inconsistent environmental conditions by transforming equirectangular images into range projections through Generative Adversarial Networks (GANs). The method uses contrastive learning to align the representations of images and point clouds. In addition, the $(LC)^2$ method (Lee et al., 2023) transforms standard images and LiDAR point clouds into disparity and range images, respectively. Consequently, the gap between the image and point cloud representations is reduced. In addition, two different architectures are employed to process disparity images and range images.

Zhao et al. (2023) developed a system that uses attention mechanisms to correlate equirectangular images and point clouds using ResNet-18 (He et al., 2016) and PointNet (Qi et al., 2017), respectively. I2P-Rec (Zheng et al., 2023) introduces a bird's eye view projection for both LiDAR point clouds and stereo-estimated point clouds. These projections are then processed using a ResNet-34 backbone for feature extraction, followed by a NetVLAD layer (Arandjelovic et al., 2016) for feature aggregation. In contrast, VXP (Li et al., 2025b) aligns the correspondences between voxels (3D units) and pixels (2D units) in a

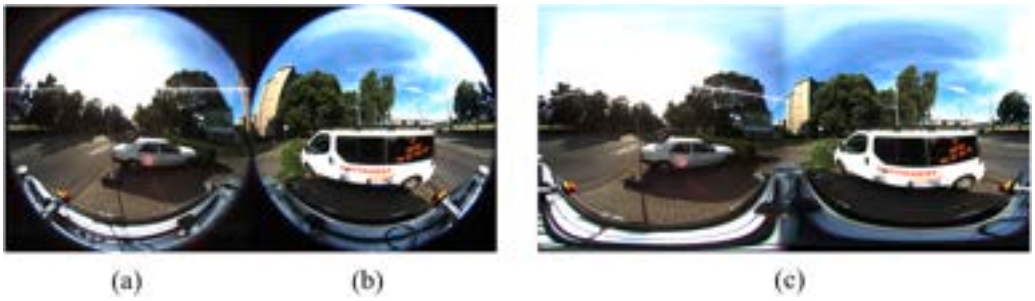

Fig. 2. Transformation of the left (a) and right (b) fisheye images to the equirectangular projection (c).

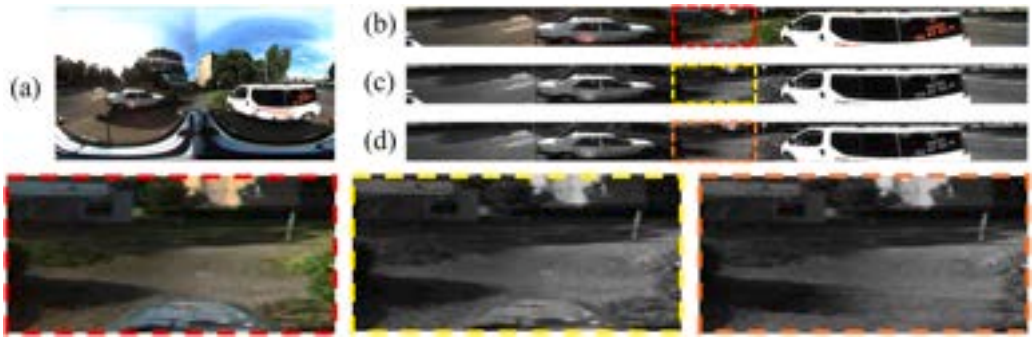

Fig. 3. Equirectangular RGB image (a); RGB image cropped to LiDAR FOV (b); grayscale image cropped to LiDAR FOV (c); and inpainting to remove occlusions from the vehicle carrying the sensors (d). Three zoomed areas are presented below, captured from the equirectangular RGB image (red), grayscale image (yellow) and the inpainted version (orange). As shown in these images, the front of the car is completely inpainted.

self-supervised manner, unifying them in a common feature space. The method uses DINO ViTs-8 (Karypidis et al., 2024) to extract features from standard images and VoxelNet (Zhou & Tuzel, 2018) for LiDAR data. This method has provided state-of-the-art results in cross-modal place recognition between standard images and LiDAR point clouds on datasets such as Oxford RobotCar (Maddern et al., 2017) and KITTI (Geiger et al., 2012).

LIP-Loc (Shubodh et al., 2024) applies the CLIP approach (Radford et al., 2021), which brings text, audio and images into the same descriptor space, but also, in this case of place recognition, between standard images and LiDAR point clouds. This method focuses on the alignment of features between standard images and point clouds projected to range images, achieving a common representation that allows cross-modal place recognition. Moreover, SaliencyI2PLoc (Li et al., 2025a) uses a dual transformer based on a ViT (Dosovitskiy et al., 2020) and PointNet (Qi et al., 2017) for cross-modal place recognition between equirectangular images and LiDAR point clouds. Moreover, Meng et al. (2025) first propose a unified learning space for LiDAR and standard cameras using a Siamese Neural Network called Cross-PRNet.

Additionally, approaches such as DistilVPR (Wang et al., 2024a) make use of distillation techniques to transfer knowledge from 3D feature extractors to image feature extractors, achieving a robust and consistent representation between modalities. Another recent approach is, VOLoc (Cai et al., 2024), which addresses the problem of embedding LiDAR data as database for efficient visual queries. Furthermore, UniLoc (Xia et al., 2024) stands out as a universal solution for urban-scale place recognition, capable of using any modality: text, image or point cloud. However, this method uses visual information to color the LiDAR point cloud, which differs from the task proposed in this paper, as our goal is to perform place recognition without combining information from different sensor modalities. Moreover, this fusion of visual and LiDAR information generally requires the calibration of both systems, which is not always straightforward (Martínez et al., 2025).

Despite advances in cross-modal place recognition, many existing approaches require training each modality network separately and then distilling the knowledge between them (Wang et al., 2024a). This can be inefficient and complicated, especially when dealing with multiple modalities. The technique proposed in this paper, CrossPlace, seeks to overcome these limitations by transforming omnidirectional views (from the two fisheye cameras) and LiDAR point clouds to a common space of intensity, depth and semantic information. As a result, a unified architecture can be employed and, furthermore, the training phase can be carried out jointly with the different sensor types.

3. Cross-modal place recognition between fisheye omnidirectional cameras and LiDAR based on a common information space

This section presents CrossPlace, the proposed approach for place recognition between omnidirectional fisheye cameras and LiDAR. The core methodology relies on bridging the modality gap by transforming raw readings from both heterogeneous sensors into a unified 2D feature space consisting of three distinct channels: intensity, depth and semantics. This pre-alignment process is specifically designed to overcome the representation conflicts and convergence bottlenecks that potentially appear in asymmetric cross-modal architectures. The extraction of these three information channels for both modalities is detailed as follows:

- **Intensity channel:** This channel captures the visual appearance and textures of the environment. For fisheye cameras, it is obtained via a direct grayscale conversion of the equirectangular RGB image. As for LiDAR, it is generated by projecting the laser remission (reflectivity) values of the point cloud into a 2D panoramic image.
- **Depth channel:** This channel provides the 3D structural geometry of the scene. For fisheye cameras, dense depth maps are inferred from the equirectangular RGB image using the pre-trained Depth Anything V2 Large model (Yang et al., 2025). In the case of the LiDAR sensor, the depth information is acquired inherently by projecting the Euclidean distance of each captured point into the panoramic space.
- **Semantic channel:** This channel captures high-level scene understanding. For fisheye cameras, semantic segmentation is performed

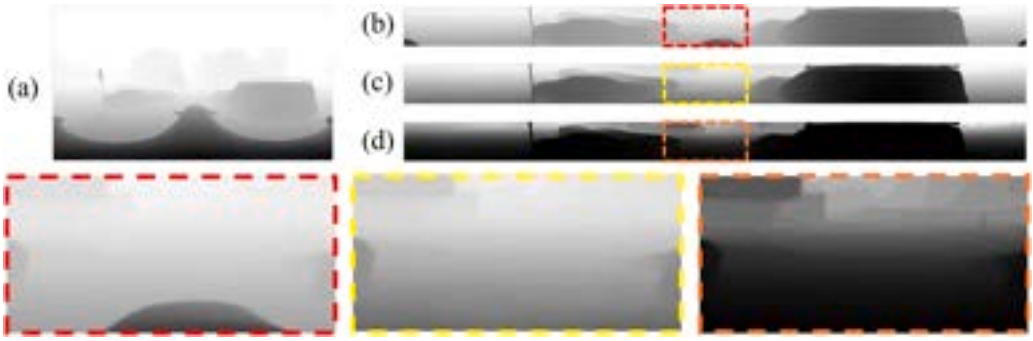

Fig. 4. Equirectangular depth image (a); cropped to LiDAR FOV (b); inpainting to remove occlusions from the vehicle carrying the sensors (c); and enhancement of the depth image to improve discrimination of distant objects (d). Three zoomed areas are presented below, captured from the equirectangular depth image (red), the inpainted version (yellow) and the inpainted and enhanced depth image (orange).

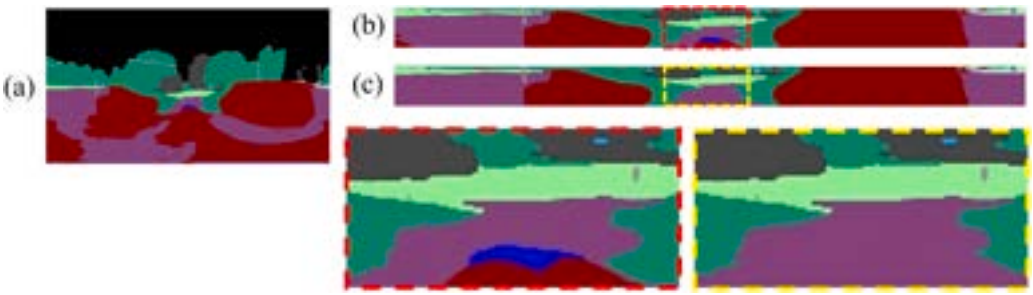

Fig. 5. Semantically segmented equirectangular image (a); cropped to LiDAR FOV (b); and inpainting to remove occlusions from the vehicle carrying the sensors (c). Two zoomed areas are presented below, captured from the equirectangular segmented image (red) and the inpainted version (yellow).

directly on the equirectangular image using the pre-trained SegFormer model (Xie et al., 2021). LiDAR data is processed by inferring point-wise semantic labels using the pre-trained MinkUNet34C model (Choy et al., 2019) and subsequently projected onto a 2D panorama. To ensure cross-modal consistency, the distinct semantic labels from both networks are harmonized into a reduced, common set of categories.

Once these three image representations are generated for a given sensor reading, they are processed by our proposed CrossPlace architecture. This framework is composed of three independent branches. Each branch is based on the CosPlace model (Berton et al., 2022) with a ResNet-152 (He et al., 2016) backbone, which has proven to be highly effective in visual place recognition.

Each branch is explicitly dedicated to processing a specific type of image: an intensity CosPlace model, a depth CosPlace model, and a semantic CosPlace model. These three models are trained independently according to their specific input domain. Finally, the embeddings extracted by each branch are fused through concatenation. This late fusion permits combining the features learned from each bridging source, which improves the ability of the method to recognize places. As demonstrated in the experiments (Section 4.5.3), different bridging sources have been evaluated independently and in combination, confirming that the fusion of intensity, depth and semantic information significantly enhances place recognition performance. By leveraging a unified preprocessing stage, CrossPlace employs the exact same network architecture and shared weights for both LiDAR and fisheye camera inputs. This unified design handles both 2D-3D and 3D-2D retrieval tasks natively, avoiding the need for distinct backbones typically required by other state-of-the-art methods to process heterogeneous data (Cattaneo et al., 2020; Li et al., 2025b; Zhao et al., 2023) (see Fig. 1).

Next, Sections 3.1 and 3.2 further detail the comprehensive data transformation and preprocessing pipelines applied to the fisheye images and LiDAR point clouds to successfully build this shared space.

3.1. Transformation of fisheye images to intensity, depth and semantic space

In the dataset used, the vision system is composed of two fisheye cameras which are rotated 180° relative to each other in order to capture opposite views of the environment. These images are then combined to generate a 360° equirectangular image. In this research, the polynomial-based geometric transformation proposed by Flores et al. (2024) is used to generate a single 360° image from the information captured by both fisheye cameras. Given a pair of fisheye images, the first step consists in converting each image into an equirectangular image. To represent the fisheye image in a spherical projection, an inverse mapping is carried out, that is: given a pixel in the equirectangular image, it is projected onto a unit sphere and then onto the fisheye image. In this way, each pixel of the final equirectangular image has an associated RGB value. To perform the projection from unit sphere to fisheye image, the camera model of Mei and Rives (2007) provided by the KITTI-360 dataset (Liao et al., 2022) is used.

Once the two equirectangular images are obtained, it is important to calculate and apply a geometric transformation to express them in the same reference system. Flores et al. (2024) propose using a polynomial to perform this geometric transformation between both equirectangular images. This is the technique used in the present paper. To estimate this transformation, correspondences between points extracted from both images are required. These correspondences are obtained through the use of ORB (Rublee et al., 2011), which computes a set of keypoints and their descriptors in both fisheye images. Subsequently, the corresponding points are used to estimate the polynomial function that best describes the transformation between the two fisheye images. Fig. 2 (a), (b) and (c) show two sample fisheye images and the resulting equirectangular image computed with the transformation process. It should be noted that the polynomial transformation model is calculated once and then applied to all fisheye image pairs in the dataset, which allows efficient and consistent conversion to equirectangular format.

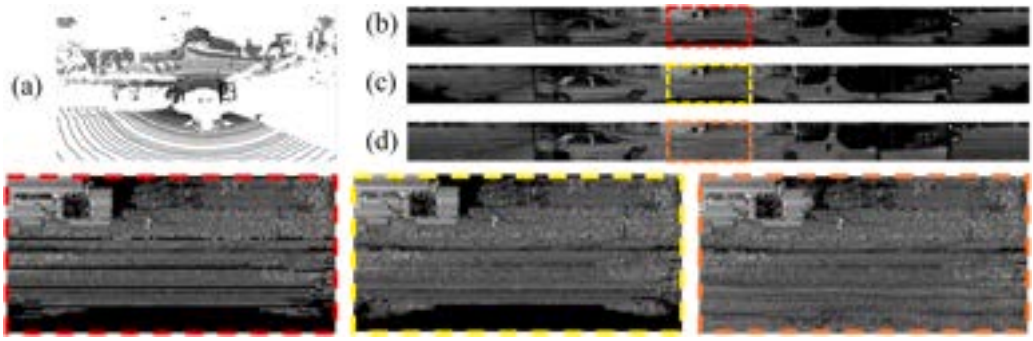

Fig. 6. Transformation process of LiDAR point clouds to intensity images. (a) LiDAR point cloud; (b) Panoramic image generated from the point cloud with LiDAR intensity; (c) Intensity panoramic image vertically interpolated to complete missing pixels; and (d) Intensity panoramic image vertically interpolated and vehicle inpainting. Three zoomed areas are presented below, captured from the LiDAR projected intensity image (red), the interpolated intensity image (yellow), and both the interpolated and inpainted version (orange).

Merging the two fisheye images into a single equirectangular image not only avoids information redundancy but also enables direct comparison with the images generated from the LiDAR sensor. By transforming both views into a single 360° image, a continuous vision of the environment, without overlaps, is obtained. Furthermore, working with a single image allows the use of a unified network architecture for both sensor modalities, which simplifies the place recognition process and improves system efficiency.

3.1.1. Transformation to intensity space

The transformation of the equirectangular image to the intensity space is directly performed through a simple conversion to grayscale (Fig. 3 (c)). This conversion is necessary to enable the comparison between equirectangular images with intensity images generated from LiDAR scans, which are also represented with one channel.

3.1.2. Transformation to depth space

Depth information is obtained from the 360° equirectangular view by Depth Anything V2 Large (Yang et al., 2025), which has been originally trained with standard images in a wide variety of scenarios and lighting conditions, making it suitable to infer distance information in complex urban environments, such as the ones used during the experiments. Therefore, this depth estimation model is used to convert the equirectangular image to a depth map (Fig. 4 (a)). The depth map provides information about the relative distance between points in the scene to the cameras, which provides relevant information for place recognition.

3.1.3. Transformation to semantic space

A third source of information is added to the place recognition architecture, with the objective of enhancing its capabilities. Specifically, the SegFormer model (Xie et al., 2021) is used to perform a semantic segmentation of the equirectangular image. Like the depth model, SegFormer has been originally trained with standard images to segment different objects in the scene, such as buildings, vehicles and pedestrians. In this paper, it is applied to each equirectangular image (Fig. 5 (a)) and the result is a image with semantic information.

Finally, the resulting intensity, depth and semantic images are cropped to match the Field Of View (FOV) of the LiDAR sensor, which permits direct comparison with point cloud scans converted to intensity, range and semantics (Figs. 3 (c), 4 (b) and 5 (b)). In addition, the LaMa inpainting model (Suvorov et al., 2021) is used to eliminate occlusions caused by the vehicle carrying the sensors (Figs. 3 (d) and 4 (c)) and, in the case of the semantically segmented image, the vehicle is removed directly by changing the semantic category of the image area where the car appears to the "road" class (Fig. 5 (c)). In addition, the depth image is transformed to achieve a greater contrast in distant areas. To do this, each pixel value (depth) is raised to the fifth power. It improves the discrimination of buildings, vegetation and other distant objects in the environment (Fig. 4 (d)).

Table 1
Parameters and values used for the spherical projection of the LiDAR point clouds.

Parameter	Value
Horizontal Field of View	360 degrees
Vertical Field of View	26.8 degrees
Projected Image Width (W)	1024 pixels
Projected Image Height (H)	64 pixels
Effective Horizontal Resolution ($\Delta\phi$)	0.35 degrees/pixel
Effective Vertical Resolution ($\Delta\theta$)	0.42 degrees/pixel

3.2. Transformation of LiDAR point clouds to intensity, depth and semantic space

LiDAR sensor point clouds (Fig. 6 (a)) are projected into a two-dimensional panoramic representation by means of a spherical projection. Given a LiDAR point cloud with coordinates (x, y, z), each point is mapped to spherical coordinates (r, ϕ, θ), where:

$$\begin{cases} r = \sqrt{x^2 + y^2 + z^2} \\ \phi = -\arctan 2(y, x) \\ \theta = \arcsin\left(\frac{z}{r}\right) \end{cases} \tag{1}$$

The angular values are discretized according to the sensor's resolution to obtain image coordinates (u, v):

$$\begin{cases} u = \left\lfloor \frac{\phi}{\Delta\phi} \right\rfloor \\ v = \left\lfloor \frac{\theta}{\Delta\theta} \right\rfloor \end{cases} \tag{2}$$

where $\Delta\phi$ and $\Delta\theta$ represent the effective angular resolution per pixel of the projected image. Note that while the raw LiDAR sensor has a specific raw horizontal resolution (approx. $0.08°$), in this work we define the effective horizontal resolution $\Delta\phi$ based on a fixed image width W to ensure consistent input dimensions for the network. Specifically, $\Delta\phi = 360°/W$. Table 1 summarizes the projection parameters used for the KITTI-360 dataset.

This process is used to generate images with intensity, depth, and semantic information, enabling direct comparison with the corresponding fisheye camera representations. The following subsections describe the transformation process for each bridging source.

3.2.1. Transformation to intensity space

To obtain the intensity image, each point cloud is projected onto a panoramic image using the spherical coordinates described above. Each

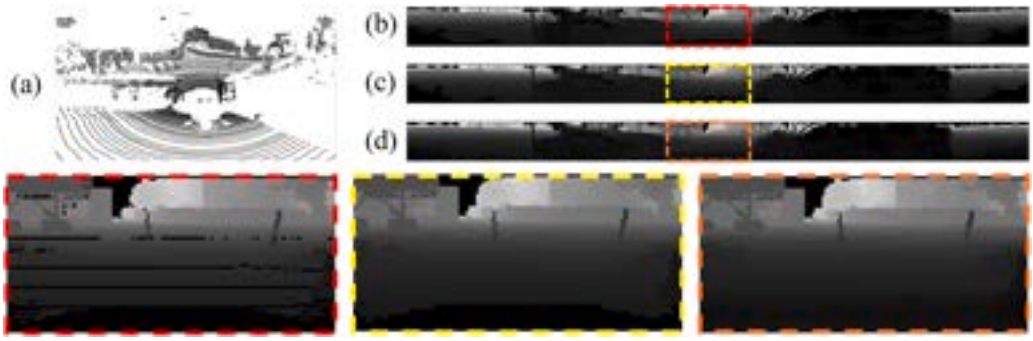

Fig. 7. Transformation process of LiDAR point clouds to range images. (a) LiDAR point cloud; (b) Panoramic image generated from the point cloud; (c) Depth panoramic image vertically interpolated to complete missing pixels; and (d) Depth panoramic image vertically interpolated and vehicle inpainting. Three zoomed areas are presented below, captured from the LiDAR projected depth image (red), the interpolated depth image (yellow), and both the interpolated and inpainted version (orange).

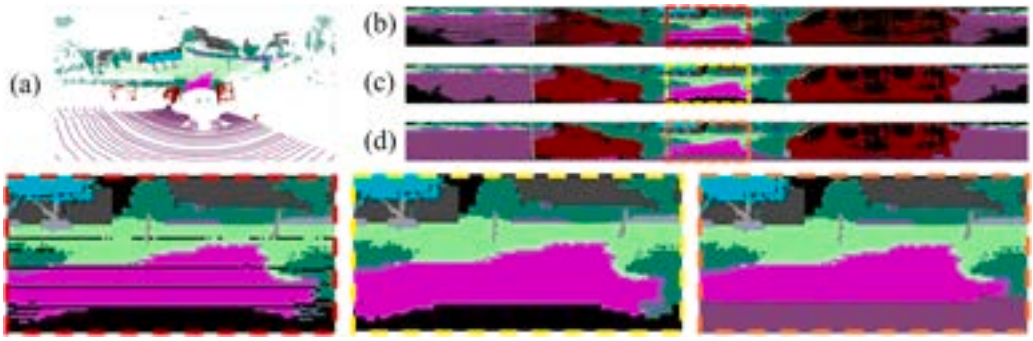

Fig. 8. Transformation process of LiDAR point clouds to semantically segmented images. (a) Segmented LiDAR point cloud; (b) Panoramic image generated from the segmented point cloud; (c) Segmented panoramic image vertically interpolated to complete missing pixels; and (d) Segmented panoramic image vertically interpolated with filtering of the vehicle carrying the sensors. Three zoomed areas are presented below, captured from the LiDAR projected segmented image (red), the interpolated segmented image (yellow), and both the interpolated and inpainted version (orange).

pixel (u, v) in the resulting image stores the intensity value of the corresponding LiDAR point, representing the laser remission (Fig. 6 (b)). The intensity reflects the amount of light returned from objects in the scene. Due to the limited vertical resolution of the LiDAR sensor, some pixels may lack data; these are completed using vertical linear interpolation (Fig. 6 (c)). Additionally, inpainting is applied to remove the vehicle carrying the sensors (Fig. 6 (d)), ensuring consistency with the fisheye camera preprocessing.

3.2.2. Transformation to depth space

For the depth image, the LiDAR point cloud is projected onto a panoramic image where each pixel (u, v) stores the Euclidean distance r from the sensor to the point (Fig. 7 (b)). Missing pixels due to vertical resolution are completed by vertical linear interpolation (Fig. 7 (c)), and inpainting is used to remove the vehicle (Fig. 7 (d)).

3.2.3. Transformation to semantic space

Regarding the semantic information of the LiDAR point cloud, it is first segmented using MinkUNet34C (Choy et al., 2019), which classifies each point in the cloud among 20 possible semantic categories (Fig. 8 (a)). Next, the segmented point cloud is projected onto a panoramic image, where each pixel (u, v) takes the color of the semantic category corresponding to the closest point in the segmented point cloud (Fig. 8 (b)). It should be noted that the semantic classes provided by MinkUNet do not exactly match the categories used by SegFormer (Xie et al., 2021). To build a common space for place recognition based on semantic information, the categories are unified into a common label set. Table 2 details the semantic harmonization strategy, mapping the specific classes from Cityscapes and SemanticKITTI to the unified categories used in CrossPlace.

Table 2
Semantic harmonization strategy. The table shows the mapping of classes from the source datasets (Cityscapes and SemanticKITTI) to the unified CrossPlace label set (ID).

ID	Unified Class	SegFormer (Cityscapes)	MinkUNet (SemanticKITTI)
0	Road	Road	Road
1	Sidewalk	Sidewalk	Sidewalk
2	Building	Building	Building
3	Wall/Fence	Wall, Fence	Fence
4	Pole/Traffic Light	Pole/Traffic Light	Pole
5	Traffic Sign	Traffic Sign	Traffic-sign
6	Vegetation	Vegetation	Vegetation, Trunk
7	Terrain	Terrain	Terrain, Other-ground
8	Person	Person	Person
9	Rider	Rider	Bicyclist, Motorcyclist
10	Car	Car	Car
11	Truck	Truck	Truck
12	Other-vehicle	Bus/Train	Other-vehicle
13	Motorcycle	Motorcycle	Motorcycle
14	Bicycle	Bicycle	Bicycle
15	Unknown	Sky	Unlabeled, Parking

Finally, there are pixels that do not have direct correspondence with LiDAR scans, due to: (1) the low vertical resolution of the LiDAR sensor, (2) points above the maximum range or below the minimum range of the sensor. To address the problem of pixels without correspondence due to the resolution of the LiDAR sensor, a vertical interpolation algorithm is applied. An example is presented in Figs. 6–8. In particular, the images in Figs. 6 (c), 7 (c) and 8 (c) are obtained from Figs. 6 (b), 7 (b) and 8 (b), respectively, by applying an interpolation on the vertical axis. This vertical interpolation algorithm relies on the heuristic

Table 3
Distribution of data pairs per sequence in the KITTI-360 dataset.

Set	Type	Sequence										
		00	02	03	04	05	06	07	08	09	10	18
Train	Query	2096	3349	-	2854	1382	2305	818	1592	3058	900	-
	Database	6917	10,320	-	8198	4909	6881	1230	5108	10,189	2126	-
Test	Query	332	-	404	-	-	-	-	-	-	-	-
	Database	1169	-	606	-	-	-	-	-	-	-	-
Further Test	Query	-	-	-	-	-	-	376	-	-	-	902
	Database	-	-	-	-	-	-	466	-	-	-	2545

Table 4
Summary of number of queries and database pairs in the KITTI-360 dataset used in the experiments, including the sequences for each environment.

Set	Urban environment				Highway environment			
	Sequences	Query	Database	Total	Sequences	Query	Database	Total
Training	*00, 02, 04, 05, 06, 07, 08, 09, 10*	18,354	55,878	74,232	-	-	-	-
Test	*00*	332	1169	1501	*03*	404	606	1010
Further Test	*18*	902	2545	3447	*07*	376	466	842

that vertically adjacent points in point clouds are likely to belong to the same object (Liu et al., 2020), a strategy also adopted by recent state-of-the-art methods like ModaLink (Xie et al., 2024) to densify sparse LiDAR representations. While this assumption implies a simplification, particularly at the boundaries of vertically overlapping structures, it effectively addresses the sparsity of LiDAR projections. Consequently, interpolated points are generated considering vertically adjacent pixels to complete the intensity, depth and semantic images. In some cases, some objects do not reflect information to the LiDAR (when the object is at a large distance or possesses a low reflectivity). In this situation, those points are assigned a zero value in the projected images (in intensity, depth and semantic information), indicating that no information is available for those pixels. Finally, points that do not have direct correspondence with LiDAR scans because they are too close to the sensor are completed through an inpainting process with LaMa (Suvorov et al., 2021) (Figs. 6 (d) and 7 (d)) and, in the case of the semantically segmented image, the pixels corresponding to the vehicle carrying the sensors are removed and assigned the "road" class (Fig. 8 (d)).

4. Experiments

4.1. Datasets

The proposed method is evaluated using the KITTI-360 dataset (Liao et al., 2022), which contains information from various sensors and has been captured in several areas of Karlsruhe (Germany), including both urban and highway environments. This dataset contains more than 320,000 images and 100,000 LiDAR scans over a distance of 73.7 km. Among the different sensors available on the vehicle, this study only uses data captured by the two fisheye cameras and the Velodyne HDL-64E LiDAR sensor. The fisheye cameras, which have a field of view of 185°, were installed on both sides of the vehicle to provide a 360° view of the environment. The LiDAR sensor is mounted on top of the vehicle and has a vertical resolution of 64 channels and a raw horizontal angular resolution of approximately 0.08°. It provides point clouds that cover a range of 120 m with a horizontal field of view of 360° and a vertical field of view of 26.8°. Additionally, it provides precise geolocated information, including GPS coordinates and vehicle orientation, which allows the labeling and identification of the sensor readings in the map.

4.2. Training and evaluation

To conduct the experiments, the dataset has been divided following the protocol established by Li et al. (2025a), splitting the different sequences that compose the dataset into urban and highway areas. Some sequences are purely urban (*00*, *02*, *04*, *05*, *06*, *08*, *09*, *10* and *18*), one sequence corresponds exclusively to highway driving (*03*), and another includes both urban and highway trajectories (*07*). Table 3 presents the distribution of data pairs per sequence in the KITTI-360 dataset, while Table 4 summarizes the total number of data pairs in each set. Note that a pair consists of an equirectangular image (constructed from the two fisheye images captured by the cameras) and a LiDAR point cloud captured at the same time instant, both serving as query or database elements depending on the evaluation mode (2D-3D or 3D-2D). For training, only the urban areas of sequences *00*, *02*, *04*, *05*, *06*, *07*, *08*, *09* and *10* were used, reserving part of sequence *00* and the highway segment of sequence 07 for testing. This training set covers a great variety of urban environments, including residential streets, commercial areas and scenarios with high traffic density. In total, 74,232 pairs of fisheye images and LiDAR point clouds were generated for training.

For evaluation, two test sets were defined, each containing both urban and highway environments. The first evaluation set, referred to as "Test", includes the urban portion of sequence *00* (different from the segment used in training) and the entire highway sequence *03*, with 1501 and 1010 data pairs, respectively (as indicated in Table 4). The second evaluation set, called "Further Test", includes the full urban sequence *18* and the highway segment of sequence *07* (not used in the training), with a total of 3447 and 842 data pairs, respectively (see Table 4). Both sets are used in the ablation study (Section 4.5) and to compare the proposed CrossPlace method with other state-of-the-art approaches (Section 4.6), enabling performance evaluation under different scenarios and conditions.

It should be noted that query data (both in training and testing) were sampled every 3 m (following (Li et al., 2025a; Zhao et al., 2023)), while the remaining readings formed the database. This ensures realistic evaluation conditions, where queries are collected at different spatial locations relative to the database readings. In total, 74,232 data pairs are available for training and 6800 data pairs for evaluation, providing a solid foundation for training and evaluation of the proposed method.

Regarding the evaluation metrics, the standard metric in place recognition has been used: recall at 1 (R@1), which measures the proportion of queries for which the closest element in the descriptor space is a true positive, i.e. it is within a distance threshold of d meters. This distance threshold d takes a value of 10 m in Section 4.5, 20 m in Section 4.6.1 and 5 m in Section 4.6.2 to ensure a fair comparison with previous research (Jiao et al., 2025; Li et al., 2025a; Shubodh et al., 2024; Xie et al., 2024; Zhao et al., 2023). The variants R@5 and R@1% will be also used in Sections 4.6.1 and 4.6.2, respectively. These metrics

are fundamental for assessing the effectiveness of the model in place recognition between different sensor modalities, given the complexity this problem poses in the current state-of-the-art.

As previously introduced, two evaluation scenarios are considered to assess cross-modal place recognition: (1) using fisheye images as queries and projected LiDAR point clouds as the database (2D-3D), and (2) using LiDAR point clouds as queries and fisheye images as the database (3D-2D). It is important to note that the proposed architecture is also capable of solving unimodal scenarios, such as 2D-2D (camera-to-camera) and 3D-3D (LiDAR-to-LiDAR) place recognition, since the unified descriptor space enables matching within the same modality. However, it is not the focus of this paper. In the 2D-3D scenario, the robot is equipped only with a pair of fisheye cameras and attempts to recognize places within a map constructed exclusively from LiDAR data. Here, query descriptors are extracted from the fisheye images using the proposed architecture and compared against database descriptors generated from LiDAR scans. Conversely, the 3D-2D scenario assumes that the robot is equipped with a LiDAR sensor and must localize itself within a map built from fisheye camera images. In this case, the query descriptor is computed from the LiDAR scan and compared with database descriptors obtained from fisheye images. Importantly, the same CrossPlace architecture is employed for both evaluation processes, ensuring consistency across modalities.

4.3. Labelling and similarity

For similarity labelling between pairs of fisheye images and LiDAR point clouds, the GPS information provided by the KITTI-360 dataset (Liao et al., 2022) is used. Thus, the similarity is defined between a query reading (whether from fisheye cameras or LiDAR) and a database element, based on the Euclidean distance between their GPS capture positions. Specifically, two captures (regardless of the sensor modality) are considered positive if they were taken within an Euclidean distance less than $p = 10$ meters from each other. Conversely, they are considered negative if the distance between them is greater than $n = 50$ meters. In this way, given a reference reading, positive and negative examples are selected independently of the sensor modality, either fisheye images or LiDAR point clouds. This criterion allows creating training pairs that promote unified learning in a single network model for both modalities, facilitating the encoding of both data types in the same descriptor space and therefore enabling cross-modal place recognition.

4.4. Implementation details

For the fine-tuning of each branch of the CrossPlace architecture, the Truncated Smooth-AP (TSAP) loss function is employed, which aims to optimize the ranking of positive candidates within the top-k results. Its formulation minimizes the gap between the smoothed average precision of each query and the ideal value, and is defined as:

$$\mathcal{L}_{TSAP} = \frac{1}{b}\sum_{q=1}^{b}(1 - AP_q) \tag{3}$$

where b denotes the batch size and AP_q corresponds to the smoothed average precision for a query reading q. The computation of AP_q is expressed as:

$$AP_q = \frac{1}{|P|}\sum_{i\in P}\frac{1+\sum_{j\in P, j\neq i} G(d(q,i)-d(q,j);\tau)}{1+\sum_{j\in\Omega, j\neq i} G(d(q,i)-d(q,j);\tau)} \tag{4}$$

where P represents the set of the k closest positives (with $k = 4$), and Ω denotes the set containing all positives and negatives. The function $G(x;\tau) = \left(1+\exp\left(-\frac{x}{\tau}\right)\right)^{-1}$ is a sigmoid parametrized by τ, which controls the ranking sharpness. The term $d(q,i)$ represents the Euclidean distance between the descriptor of the query point cloud (or image) q and the i-th image (or point cloud). The numerator expresses a soft ranking of a positive point i among the top-k positives, while the denominator evaluates its ranking among all the other candidates.

Table 5
Parameters and values used to fine-tune each branch of the CrossPlace architecture.

Parameter	Value
Positive distance (p)	10 m
Negative distance (n)	50 m
Batch size	1024
Number of epochs	4
Initial Learning Rate (LR)	1×10^{-3}
LR Scheduler Steps	2, 3
L2 Weight Decay	1×10^{-4}
Positives per Query (k)	4
Sigmoid sharpness (τ)	0.01
Distance threshold (d)	10 m

For effective performance, this loss function requires a large batch size (Komorowski, 2022). In this approach, a batch size of 1024 elements (images or point clouds) is selected, which ensures better convergence and training stability. The model is trained for 4 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3}. The learning rate is reduced by a factor of 10 at epochs 2 and 3, allowing for finer convergence at the later training stages. The parameters employed to train CrossPlace are summarized in Table 5.

All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU with 24 GB. Our code is publicly available on the project website.[1]

4.5. Ablation study

In this section, the influence of the most relevant parts of the framework is analysed. Specifically, Section 4.5.1 studies the performance of the proposed method depending on the bridging source of information used to compare fisheye images and LiDAR point clouds (i.e. intensity, depth of semantic information). Subsequently, Section 4.5.2 analyses the impact of preprocessing intensity, depth and semantic images, assessing the influence of the different preprocessing techniques on the quality of the extracted features. Finally, in Section 4.5.3, different fusion techniques of features extracted from intensity, depth and semantic images are studied to determine their influence on cross-modal place recognition.

4.5.1. Intensity, depth and semantics as bridging sources between LiDAR and camera

In this section, the influence of each bridging source (intensity, depth and semantics) is studied, evaluating each of them separately to clarify their individual contributions to cross-modal place recognition. Specifically, each branch of the CrossPlace architecture is trained and tested independently (without preprocessing), allowing a direct comparison of their effectiveness in bridging the camera and LiDAR modalities. For this purpose, each branch is trained and evaluated separately in terms of R@1 with the Test and Further Test sets, differentiating between urban and highway environments, and between 2D-3D and 3D-2D modalities. As described in Section 3, intensity information is obtained directly from the equirectangular images provided by the fisheye cameras, and from the LiDAR intensity. Additionally, depth information is also obtained directly from LiDAR projection, but for depth estimation in equirectangular images Depth Anything V2 Large (Yang et al., 2025) is used. Finally, semantic information is obtained from equirectangular images using the SegFormer (Xie et al., 2021) model and from LiDAR point cloud using the MinkUNet34C (Choy et al., 2019) model.

The results shown in Table 6 indicate that intensity information presents the lowest performance in all cases, especially on highways, where the R@1 does not exceed 76% in any of the modalities. However, as indicated previously, it is the only bridging source that does

[1] https://juanjo-cabrera.github.io/projects-CrossPlace/

Table 6
An independent evaluation of the different bridging sources (intensity, depth and semantics) between LiDAR and fisheye cameras. The table shows the R@1 (expressed as a percentage) for a distance $d = 10m$ on the Test and Further Test sets, differentiating between urban and highway environments, and between 2D-3D and 3D-2D queries. The last column shows the total average of the 8 values.

	Test				**Further Test**				
	Urban (00)		Highway (03)		Urban (18)		Highway (07)		
Source	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	**Total**
Intensity	98.19	97.89	75.25	72.03	94.12	94.24	63.30	62.23	82.16
Depth	98.80	98.19	**87.62**	**88.61**	95.57	95.23	**82.18**	**80.32**	**90.82**
Semantics	**99.70**	**100.00**	78.22	85.40	**98.67**	**97.23**	76.06	72.12	88.68

Table 7
The influence of the different preprocessing techniques applied to the grayscale image (2D) and the LiDAR intensity (3D) on the performance of the CrossPlace intensity branch. The table shows the R@1 (expressed as a percentage) for a distance $d = 10m$ on the Test and Further Test sets, differentiating between urban and highway environments, and between 2D-3D and 3D-2D queries. The last column shows the total average of the 8 values.

		Test				**Further Test**				
		Urban (00)		Highway (03)		Urban (18)		Highway (07)		
	Intensity	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	**Total**
	Baseline	98.19	97.89	75.25	72.03	94.12	**94.24**	63.30	62.23	82.16
2D	+ Static Mask [a]	98.34	98.22	76.15	73.12	93.45	92.88	64.81	61.67	82.33
	+ inpainting [b]	**98.49**	**98.49**	**77.23**	74.50	92.79	91.80	66.76	60.90	82.62
3D	+ Interpolation	98.19	97.89	76.98	**76.49**	**96.34**	92.57	68.88	**72.87**	**85.03**
	+ inpainting [c]	97.89	98.19	73.02	72.28	94.12	90.24	**71.81**	69.68	83.40

[a] Application of a static zero-cost mask over the ego-vehicle region.
[b] Inpainting the ego-vehicle region in the grayscale equirectangular image.
[c] Inpainting the ego-vehicle region in the LiDAR intensity image.

not require any prior learning model, since it is based solely on the intensity of LiDAR scans and grayscale images from cameras. Regarding semantic information as a bridging source, it is the most robust in urban environments reaching a R@1 close to 100% in most cases. However, on highways, this type of information presents poorer performance. Specifically, if we compare the results of semantic information with depth in highway environments, depth surpasses semantics in all cases, being a more suitable bridging source for recognition in more homogeneous and repetitive environments like highways, where scene geometry is crucial for place identification. In general, depth information shows superior performance compared to intensity and semantics in all cases, reaching an average R@1 of 90.82%, but when it comes to urban environments, semantics surpasses depth information.

4.5.2. Preprocessing of intensity, depth and semantics

In this section, the influence of preprocessing intensity, depth and semantic images before feeding each branch of the CrossPlace architecture is studied independently. The results obtained will demonstrate that the preprocessing stage is crucial to improve the cross-modal retrieval in the database using both sensor modalities. Specifically, two preprocessing techniques are evaluated, as described in Section 3: (1) inpainting the vehicle carrying the sensors in both fisheye and LiDAR images and (2) vertical interpolation of LiDAR images.

Table 7 shows the R@1 results of the intensity branch with different preprocessing techniques on both LiDAR intensity images and grayscale equirectangular images. First, the baseline result is obtained without any preprocessing (raw data), achieving an average R@1 = 82.16%. Next, to evaluate a computationally efficient alternative for handling ego-vehicle occlusion, a zero-cost static mask is applied over the occluded region. This simple masking slightly improves the baseline to 82.33%, demonstrating that the network can learn to partially ignore this area without additional processing overhead. However, when an advanced inpainting technique is applied to actively reconstruct the road texture from the grayscale equirectangular image, the global performance further increases to 82.62%. This indicates that the framework is robust to potential artifacts generated by the inpainting process. The removal of the vehicle from the images provides benefits that outweigh the minor artifacts in the reconstructed road texture. From now on, this inpainting technique will be applied to the 2D intensity images. Subsequently, the vertical interpolation technique is applied to the LiDAR intensity image, which further improves the R@1 to 85.03%. This technique is especially effective in highway sequences and specifically in the 3D-2D modality, where the query images come from LiDAR and the database consists of equirectangular fisheye views, reaching 76.49% and 72.87% in sequences *03* and *07*, respectively (from a baseline R@1 of 72.03% and 62.23%). Finally, in addition to the vertical interpolation, inpainting is applied to the LiDAR intensity image to remove the vehicle carrying the sensors. However, this processing on the input LiDAR intensity image does not produce the desired effect on the global performance of the intensity branch. Therefore, from here on, only both the vertical interpolation for LiDAR intensity images and the inpainting for grayscale equirectangular images will be used.

Subsequently, Table 8 presents the results of the depth branch when preprocessing techniques are applied to the equirectangular depth images (obtained through Depth Anything V2) and the spherical projection of LiDAR. To begin with, the baseline does not have any preprocessing, obtaining an average R@1 of 90.82%. Similar to the intensity branch, applying a zero-cost static mask over the ego-vehicle yields a noticeable improvement (91.26%). Nevertheless, employing the advanced inpainting technique to actively reconstruct this region in the 2D equirectangular depth image produces a superior global improvement of 1.05% (R@1 = 91.87%). This improvement is especially notable in urban environments, where recall values of 99.70% and 100.00% are reached in the sequence *00* and values of 98.12% and 98.34% in the sequence *18*,

Table 8
The influence of the different preprocessing techniques applied to the depth image (2D) and the LiDAR range image (3D) on the performance of the CrossPlace depth branch. The table shows the R@1 (expressed as a percentage) for a distance $d = 10m$ on the Test and Further Test sets, differentiating between urban and highway environments, and between 2D-3D and 3D-2D queries. The last column shows the total average of the 8 values.

		Test				**Further Test**				
		Urban (00)		Highway (03)		Urban (18)		Highway (07)		
	Depth	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	**Total**
	Baseline	98.80	98.19	87.62	88.61	95.57	95.23	82.18	80.32	90.82
2D	+ Static Mask [a]	99.12	99.04	87.75	89.84	96.81	96.75	81.25	79.52	91.26
	+ inpainting [b]	**99.70**	**100.00**	87.87	91.09	98.12	98.34	80.32	79.52	91.87
	+ Equalization	100.00	99.70	88.37	90.35	98.89	98.89	82.71	77.66	92.07
	+ Power ($p = 2$)	100.00	99.70	91.83	**94.06**	98.89	99.45	84.31	75.53	92.97
	+ Power ($p = 3$)	99.40	99.70	94.31	91.00	98.23	99.11	84.04	80.59	93.19
	+ Power ($p = 4$)	100.00	99.70	90.03	92.57	98.89	98.67	84.04	82.04	93.24
	+ Power ($p = 5$)	99.40	99.70	92.08	**94.06**	98.67	99.00	83.51	82.18	93.57
	+ Power ($p = 6$)	99.40	99.70	90.03	91.34	98.52	98.96	83.07	80.31	92.67
3D	+ Interpolation	**99.70**	99.40	**95.54**	92.08	**99.22**	**99.22**	**88.03**	**85.37**	**94.82**
	+ inpainting [c]	99.40	99.70	92.08	93.32	98.89	99.11	85.11	82.45	93.76

[a] Application of a static zero-cost mask over the ego-vehicle region.
[b] Removal of the vehicle on which the sensors are mounted in the equirectangular depth image.
[c] Removal of the vehicle on which the sensors are mounted in the LiDAR range image.

Table 9
The influence of the different preprocessing techniques applied to the segmented image (2D) and the segmented LiDAR (3D) on the performance of the CrossPlace semantic branch. The table shows the R@1 (expressed as a percentage) for a distance $d = 10m$ on the Test and Further Test sets, differentiating between urban and highway environments, and between 2D-3D and 3D-2D queries. The last column shows the total average of the 8 values.

		Test				**Further Test**				
		Urban (00)		Highway (03)		Urban (18)		Highway (07)		
	Semantics	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	**Total**
	Baseline	99.70	**100.00**	78.22	85.40	98.67	97.23	76.06	72.12	88.68
2D	+ inpainting [a]	**100.00**	**100.00**	80.20	84.16	97.01	97.12	78.19	72.87	88.69
3D	+ Interpolation	98.49	99.70	80.20	85.45	**99.00**	97.78	**80.05**	**76.06**	89.59
	+ inpainting [b]	98.80	**100.00**	**83.17**	**87.87**	98.34	**98.23**	79.79	75.80	**90.25**

[a] Removal of the vehicle on which the sensors are mounted in the equirectangular semantic image.
[b] Removal of the vehicle on which the sensors are mounted in the LiDAR semantic image.

respectively. This result further confirms that the method is robust to inpainting artifacts, as the geometric continuity provided by the removal of the vehicle contributes more to the descriptor distinctiveness than the presence of the occlusion. Regarding contrast enhancement, although Histogram Equalization improves over the baseline (92.07%), the fifth power transformation yields significantly better results (93.57%). This technique is especially effective in the highway sequence *03*, where a R@1 of 94.06% is reached in the 3D-2D modality. From now on, both the inpainting and powering techniques will be used for processing equirectangular depth images. Regarding the processing of the 3D LiDAR range image, the vertical interpolation technique is applied, which also boosts the global performance to 94.82%. Finally, inpainting the interpolated 3D image to remove the vehicle carrying the sensors does not have the desired contribution, as in the case of intensity. Therefore, only the vertical interpolation will be applied to preprocess LiDAR range images and both inpainting and powering will be used to preprocess equirectangular depth images.

Regarding the processing of semantically segmented images, Table 9 shows the R@1 results of the semantic branch when different preprocessing techniques are applied to the segmented images. First, the baseline (i.e. no preprocessing is applied to the semantic images) presents a global R@1 of 88.68%. Next, the inpainting technique is applied to the segmented equirectangular image, removing the vehicle carrying the sensors, which produces a slight improvement of 0.01% (R@1 = 88.69%). From now on, this inpainting technique will be applied to 2D semantic images. Subsequently, the vertical interpolation technique applied to the segmented LiDAR images further improves the global performance to 89.59%. This technique is especially effective in the highway sequence *07*, where a recall of 80.05% is reached in the 2D-3D modality and 76.06% in the 3D-2D modality. Finally, inpainting is applied to the segmented LiDAR image, which significantly improves the global performance to 90.25%. For this reason, from here on, both the vertical interpolation and the inpainting techniques will be used for segmented LiDAR images, and the inpainting will also be applied to segmented equirectangular images.

4.5.3. Early vs. late fusion of the bridging sources of information

In this section, the impact of early and late fusion of the different bridging sources of information on the performance of the cross-modal place recognition task is studied. First, starting either from an equirectangular image or from a LiDAR point cloud, the intensity image (1 channel), the depth image (1 channel) and the semantic information image (3 channels) are obtained. Early fusion consists of concatenating these three images along the channel dimension before being fed into the model. This results in a single input image with 5 channels, requiring the adaptation of the first convolutional layer of the network to accept 5 input channels. It is important to note that, with early fusion, the architecture only has a single branch that processes the fused input. In contrast, late fusion uses an independent branch for each kind of bridging source and then combines the embeddings extracted by each

Table 10
Evaluation of different fusion techniques for intensity, depth and semantic bridging sources. The table shows the R@1 (expressed as a percentage) for a distance $d = 10$ m on the Test and Further Test sets, differentiating between urban and highway environments, and between 2D-3D and 3D-2D queries. The last column shows the total average of the 8 values.

	Test				**Further Test**				
	Urban (00)		Highway (03)		Urban (18)		Highway (07)		
Source	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	**Total**
Intensity	98.19	97.89	76.98	76.49	96.34	92.57	68.88	72.87	85.03
Depth	99.70	99.40	95.54	92.08	99.22	99.22	88.03	85.37	94.82
Semantics	98.80	**100.00**	83.17	87.87	98.34	98.23	79.79	75.80	90.25
Early fusion	99.70	99.70	90.59	93.81	99.11	98.78	83.51	78.19	92.94
Late fusion [a]	**100.00**	**100.00**	95.30	96.53	**99.89**	99.67	93.88	90.43	96.96
Late fusion [b]	**100.00**	**100.00**	**96.29**	**98.02**	99.78	**99.89**	**94.41**	**91.22**	**97.45**

[a] Addition of intensity, depth and semantic descriptors.
[b] Concatenation of intensity, depth and semantic descriptors.

branch. Late fusion is evaluated in two ways: (1) by adding the intensity, depth, and semantic embedding descriptors, and (2) by concatenating these descriptors.

Table 10 shows the R@1 results for different bridging sources (intensity, depth and semantics) after processing each of them, both individually and combined (considering either early or late fusion). The 2D-3D and 3D-2D modalities are evaluated again on the Test and Further Test sets, differentiating between urban and highway environments. As described previously, when using only the intensity branch, the performance of the method takes an average R@1 of 85.03%, whereas using uniquely depth or semantic information the model reaches a recall of 94.82% and 90.25%, respectively. Regarding early fusion, the global R@1 shows that the framework does not take advantage of the three sources of information, reaching a value of 92.94%, although it surpasses the individual performance of the semantic and intensity branches. This suboptimal performance can be explained by the concept of the modality dominance phenomenon (Liu et al., 2026). Specifically, in early fusion, where intensity, depth and semantics are stacked as input channels, the network jointly optimizes the initial filters. Because these domains possess different statistical properties and convergence rates, the loss function tends to greedily favor the modality that minimizes the error most rapidly. This behavior effectively suppresses the gradients of the other sources of information, preventing the network from learning truly complementary features.

The late fusion of the output of the three branches consistently outperforms both the individual sources and their early fusion. This phenomenon stems from the fact that late fusion completely decouples the representation learning phase. By processing each source through independent branches, the architecture forces the network to extract the maximum discriminative features from each domain. This ensures an equitable and robust contribution from visual textures, 3D geometry and scene understanding in an independent manner (without initially merging them). Specifically, the additive fusion of intensity, depth and semantic descriptors significantly improves the global performance to 96.96%. However, the concatenation of these embeddings achieves the best global performance with an outstanding 97.45%, surpassing all other modalities and bridging sources. In particular, it stands out especially in highway sequences, which are repetitive and challenging scenarios.

The superiority of concatenation over additive fusion can be explained analytically: while additive fusion merges independent embeddings element-wise (forcing fundamentally different concepts to collide in the same latent space and rendering the representation susceptible to destructive interference), concatenation orthogonally preserves the distinct dimensionality of each feature space. This effectively maintains the complementary nature of the cross-modal information. These observations align with findings in previous state-of-the-art multi-modal place recognition works, such as MinkLoc++ (Komorowski et al., 2021), PIC-Net (Lu et al., 2020) and AdaFusion (Lai et al., 2022).

Therefore, it can be concluded that a late fusion through concatenation of intensity, depth and semantic information is the most suitable option for cross-modal place recognition between fisheye cameras and LiDAR. CrossPlace is finally defined as a three-branch architecture for cross-modal place recognition between fisheye omnidirectional cameras and LiDAR, using late fusion through concatenation of intensity, depth and semantic embeddings.

4.6. Comparison with the state-of-the-art

In this section, the proposed CrossPlace architecture is comprehensively evaluated against current state-of-the-art methods. Specifically, Section 4.6.1 presents the comparative results on the standard KITTI-360 benchmark, analyzing the performance under various strict distance thresholds to demonstrate its metric localization precision. Subsequently, Section 4.6.2 assesses the adaptability and robustness of the method on the NCLT dataset, validating its capability to generalize across different sensor configurations and severe long-term seasonal variations.

4.6.1. Evaluation on the KITTI-360 dataset

In this section, CrossPlace is compared with state-of-the-art methods that also use the KITTI-360 dataset (Liao et al., 2022) to perform cross-modal place recognition between fisheye cameras and LiDAR. Specifically, the results obtained are compared with those of AE-Spherical (Zhao et al., 2023), LIP-Loc (Shubodh et al., 2024) and SaliencyI2PLoc (Li et al., 2025a). In general, it is difficult to compare the method with all the state-of-the-art approaches, but the authors of SaliencyI2PLoc (Li et al., 2025a) tried to unify the different solutions when it comes to the KITTI-360, defining a new comparison benchmark. For this purpose, they trained the LIP-Loc and AE-Spherical models following the benchmark that they established in their own approach, and evaluated them under the same conditions. For this reason, CrossPlace has been trained and evaluated following the exact same KITTI-360 dataset division proposed by Li et al. (2025a).

It should be noted that these prior studies use a distance tolerance d of 20 m to evaluate the performance of their models. While we report our results at this standard $d = 20$ m threshold to ensure a fair comparison, this often leads to near-saturation (results close to 100%) in urban sequences. To ensure that the evaluation protocol remains sufficiently challenging and to analyze the true metric precision of our model, we propose evaluating our model under significantly stricter distance thresholds ($d = 1$ m, 2 m, 5 m and 10 m).

Table 11 presents the CrossPlace results across these varied thresholds. As expected, the near-saturation observed at 20 m and 10 m is

Table 11
CrossPlace results in terms of R@1 for varying distance thresholds from $d = 1$ m to $d = 20$ m. Results are shown for 2D-3D and 3D-2D modalities on the Test and Further Test sets, differentiating between urban and highway environments. The last column shows the total average of the 8 values.

	Test				**Further Test**				
	Urban (00)		Highway (03)		Urban (18)		Highway (07)		
Distance d	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D	**Total**
CrossPlace 1 m	66.24	58.52	50.00	36.96	35.40	33.78	57.14	42.86	47.61
CrossPlace 2 m	91.27	87.35	70.54	72.77	72.73	68.96	60.85	60.85	73.17
CrossPlace 5 m	99.70	99.10	85.64	87.62	98.34	96.56	80.05	73.94	90.12
CrossPlace 10 m	100.00	100.00	96.29	98.02	99.78	99.89	94.41	91.22	97.45
CrossPlace 20 m	100.00	100.00	99.50	99.50	100.00	100.00	98.67	96.54	99.28

Table 12
CrossPlace results compared to the state-of-the-art in KITTI-360. The table shows the Recall@1 and Recall@5 (expressed as a percentage) for the standard distance $d = 20$ m threshold. To isolate the architectural contribution from semantic and depth priors, we present the baseline SaliencyI2PLoc [‡] (Li et al., 2025a) performance fed with our multi-modal inputs, as well as our architecture using only unaugmented sensor representations (*Intensity branch*).

		Test				**Further Test**				
		Urban (00)		Highway (03)		Urban (18)		Highway (07)		**Total**
	Method	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	**R@1**
2D-3D	AE-Spherical (Zhao et al., 2023)	41.57	60.54	10.64	22.28	37.14	60.42	4.79	20.74	23.54
	LIP-Loc (Shubodh et al., 2024)	64.46	79.52	37.87	58.42	-	-	-	-	-
	SaliencyI2PLoc (Li et al., 2025a)	78.92	86.75	30.94	49.26	61.86	81.15	22.34	50.27	48.52
	SaliencyI2PLoc[a] (Li et al., 2025a)	67.47	81.93	36.39	56.19	41.24	62.97	23.67	38.83	42.19
	SaliencyI2PLoc[b] (Li et al., 2025a) (+ FM)	63.25	84.64	33.42	53.96	41.46	68.07	28.72	43.35	41.71
	CrossPlace (Intensity branch)	99.70	100.00	85.89	98.02	98.23	99.78	81.38	95.48	91.30
	CrossPlace (ours)	**100.00**	**100.00**	**99.50**	**99.75**	**100.00**	**100.00**	**98.67**	**100.00**	**99.54**
3D-2D	SaliencyI2PLoc[a] (Li et al., 2025a)	67.47	84.64	34.16	54.70	44.57	66.52	23.40	43.88	42.40
	SaliencyI2PLoc[b] (Li et al., 2025a) (+ FM)	60.54	85.54	32.18	57.92	43.90	70.95	23.67	46.01	40.07
	CrossPlace (Intensity branch)	98.49	100.00	88.37	97.52	95.90	99.33	82.18	93.88	91.24
	CrossPlace (ours)	**100.00**	**100.00**	**99.50**	**100.00**	**100.00**	**100.00**	**96.54**	**100.00**	**99.01**

[a] Results obtained by retraining the official open-source implementation.
[b] Results of SaliencyI2PLoc (Li et al., 2025a) retrained using the same Foundation Models inputs (Depth Anything V2 Yang et al., 2025, SegFormer Xie et al., 2021, MinkUNet Choy et al., 2019) as CrossPlace.

effectively mitigated when applying stricter tolerances. At a highly restrictive distance of $d = 1$ m, the global average performance drops to 47.61%, revealing the inherent difficulty of the benchmark for exact metric localization, particularly in visually aliased highway environments (e.g., 36.96% in sequence 03 for 3D-2D). Nevertheless, at a still highly strict threshold of $d = 5$ m, CrossPlace demonstrates exceptional robustness, maintaining a global average R@1 of 90.12%.

Next, Table 12 shows the comparison of CrossPlace against the state-of-the-art methods in terms of R@1 for the standard 20 m threshold. It is important to note that these prior approaches only consider the 2D-3D modality (using images to query LiDAR databases). To provide a complete overview, our results are reported for both 2D-3D and 3D-2D tasks. Furthermore, to establish a fair and direct comparison, we retrained the official open-source implementation of the best-performing baseline, SaliencyI2PLoc (Li et al., 2025a), under our exact evaluation framework (denoted as SaliencyI2PLoc[†]).

To rigorously isolate the architectural contributions of CrossPlace from the high-quality priors provided by Foundation Models (FMs), we conducted a two-fold comparative analysis. First, we retrained SaliencyI2PLoc (Li et al., 2025a) feeding it the exact same multi-modal inputs used in our pipeline, including depth and semantic maps (denoted as SaliencyI2PLoc[‡]). Feeding these rich representations into the baseline did not translate into performance gains, with the overall R@1 remaining almost unchanged at 41.71% in the 2D-3D direction. Second, we evaluated the core capacity of our architecture without any FM priors via the *Intensity branch*, which strictly uses unaugmented sensor inputs (grayscale images and LiDAR remission). Remarkably, even in this constrained setup, CrossPlace achieves an average R@1 of over 91% in both directions, significantly outperforming the baseline even when the latter is fed with high-quality depth and semantic maps.

In general, CrossPlace surpasses all state-of-the-art approaches across all sequences. Specifically, it stands out in highway sequences, where proposals such as AE-Spherical (Zhao et al., 2023), LIP-Loc (Shubodh et al., 2024) and SaliencyI2PLoc (Li et al., 2025a) suffer from severe performance drops, achieving R@1 values of only 10.64%, 37.87%, and 30.94%, respectively, in sequence *03*. In stark contrast, CrossPlace achieves 99.50% in the same sequence. In urban environments, CrossPlace also outperforms existing methods, saturating the R@1 metric to 100.00% in sequences *00* and *18*. Overall, CrossPlace achieves a global average performance of 99.54% (2D-3D) and 99.01% (3D-2D), significantly pushing the boundaries of the established benchmark and proving to be the most effective method for cross-modal place recognition on the KITTI-360 dataset.

Finally, to ensure that the performance gains of the proposed architecture over existing methods are robust, we evaluated the statistical significance of these results. Since the evaluation relies on paired nominal data (query-by-query success or failure at R@1), we applied a McNemar's test to compare the predictions of CrossPlace against the best-performing baseline, SaliencyI2PLoc (Li et al., 2025a). In a direct paired comparison across the evaluation sets, CrossPlace successfully retrieved 50.3% of the queries where the baseline failed, whereas the opposite occurred in only 1.9% of the instances. The statistical test yielded a highly significant p-value ($p \ll 0.001$) across all individual test sequences as well as globally. This confirms that the superior retrieval accuracy achieved by CrossPlace is statistically significant, definitively validating the effectiveness of the proposed unified descriptor space.

Table 13
CrossPlace results compared to state-of-the-art methods on the NCLT dataset. The method demonstrates strong generalization across long-term seasonal changes and different sensor characteristics (Velodyne HDL-32E). The final column shows the average R@1 across the four query sequences with a distance threshold $d = 5$ m.

		2012-02-05		2012-06-15		2013-02-23		2013-04-05		**Total**
	Method	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	**R@1**
2D-3D	ModaLink (Xie et al., 2024)	63.80	89.90	47.20	81.20	33.90	76.30	35.90	72.40	45.20
	InsCMPR (Jiao et al., 2025)	71.30	94.90	49.30	81.40	41.20	82.00	41.30	76.30	50.78
	CrossPlace (ours)	**91.98**	**99.00**	**72.33**	**92.73**	**78.09**	**96.62**	**75.87**	**93.61**	**79.57**
3D-2D	ModaLink (Xie et al., 2024)	61.10	87.40	53.50	83.80	36.30	72.90	36.10	70.40	46.75
	InsCMPR (Jiao et al., 2025)	71.20	95.10	62.40	88.70	45.70	83.80	42.80	75.00	55.53
	CrossPlace (ours)	**94.88**	**99.52**	**87.01**	**97.34**	**65.84**	**91.92**	**62.48**	**88.89**	**77.55**

4.6.2. Evaluation on the NCLT dataset

To assess the generalization capability of CrossPlace across different sensor configurations and environments, we conducted additional experiments on the North Campus Long-Term (NCLT) dataset (Carlevaris-Bianco et al., 2016). Unlike KITTI-360, NCLT captures a university campus environment featuring diverse structural elements and dynamic objects. Crucially, the sensor setup differs significantly, employing a Ladybug3 omnidirectional camera and a Velodyne HDL-32E LiDAR. The latter provides only 32 vertical beams (compared to 64 in KITTI-360), allowing us to evaluate the robustness of the method when dealing with lower-resolution sparse geometry.

Following the protocol established in prior works (Jiao et al., 2025; Xie et al., 2024), the sequence captured on *2012-01-08* is used to build up the database. The evaluation is performed on four query sequences (*2012-02-05*, *2012-06-15*, *2013-02-23* and *2013-04-05*) that encompass a 15-month period, capturing pronounced long-term variations.

Table 13 summarizes the comparative results against the state-of-the-art methods ModaLink (Xie et al., 2024) and InsCMPR (Jiao et al., 2025). In the 2D-3D task, CrossPlace demonstrates remarkable robustness, outperforming the closest competitor (InsCMPR Jiao et al., 2025) by large margins. Our method achieves an overall average R@1 of 79.57%, well above the 50.78% obtained by InsCMPR. For instance, in the particularly challenging *2013-02-23* sequence, our method achieves a R@1 of 78.09%, compared to 41.20% for InsCMPR (Jiao et al., 2025). This result indicates that our approach effectively bridges the domain gap in a variety of environments, even when the characteristics of the sensors change and in presence of seasonal variations.

In the 3D-2D task, despite the significantly sparser LiDAR input (32 beams compared to 64 in KITTI-360), CrossPlace demonstrates high robustness, consistently maintaining superior performance across all seasonal changes with an average R@1 of 77.55%. Notably, in the *2012-02-05* sequence, CrossPlace achieves 94.88% in R@1, surpassing InsCMPR (Jiao et al., 2025) by over 23 percentage points. These results confirm that the proposed method generalizes effectively to new environments and sensor setups.

4.7. System analysis and discussion

This section provides a deeper analysis of the overall performance and practical viability of the CrossPlace system. First, Section 4.7.1 examines the Precision-Recall (PR) curves and metric localization errors to better understand the system's robustness, highlighting characteristic challenges such as longitudinal ambiguity in highway environments. Then, Section 4.7.2 details the computational efficiency and runtime of the proposed architecture, confirming its suitability for real-world robotic applications. Finally, Section 4.7.3 presents qualitative results, providing visual examples of successful cross-modal retrievals and examining the most complex scenarios.

4.7.1. Precision-recall and localization error analysis

To further evaluate the robustness of CrossPlace, we analyze the Precision-Recall (PR) curves (using a distance threshold of 5 meters to define a true positive) and the metric localization error of the retrieved candidates. Fig. 9 illustrates these metrics for representative urban and highway sequences.

As shown in Fig. 9 (a), the method achieves near-perfect performance in urban scenarios (Sequences *00* and *18*), with Average Precision (AP) scores exceeding 0.97 for both 2D-3D and 3D-2D modalities. This confirms that the learned descriptors are highly discriminative in structurally rich environments. In contrast, highway sequences (*03* and *07*) exhibit a drop in AP (ranging from 0.72 to 0.87). This behavior is expected due to the high degree of visual aliasing and repetitive structures that characterize highway environments.

Fig. 9 (b) presents the average distance error of the Top-1 retrieved candidate. Consistent with the PR analysis, urban sequences show low localization errors (1.21 m – 1.70 m), indicating that the retrieved matches are geometrically very close to the query. Conversely, highway sequences show higher errors (2.98 m – 4.67 m). This increase is attributed to the higher ambiguity of highways, where the retrieved frame often corresponds to a visually identical road segment located a few meters ahead or behind the true position. Nevertheless, the error remains within acceptable bounds for a coarse localization step in outdoor scenarios.

Table 14
Breakdown of inference time per frame for 2D-3D and 3D-2D tasks on an RTX 3090 GPU.

Task (Query)	**Module**	**Time (ms)**	**Freq (Hz)**
2D-3D (Camera)	Semantic Seg. (SegFormer)	142.12	
	Depth Est. (Depth Anything V2)	263.83	
	Inpainting (LaMa)	35.32 (17.66 × 2)	
	Feature Extraction	15.03 (5.01 × 3)	
	Total Latency	**456.30 ms**	**≈ 2.19 Hz**
3D-2D (LiDAR)	Semantic Seg. (MinkUNet34C)	79.76	
	Spherical Projection	9.55	
	Vertical Interpolation	0.51 (0.17 × 3)	
	Feature Extraction	15.03 (5.01 × 3)	
	Total Latency	**104.85 ms**	**≈ 9.54 Hz**

4.7.2. Runtime analysis

In real-world robotic applications, inference speed is a critical factor alongside retrieval accuracy. We evaluated the runtime performance of the proposed method on a desktop computer equipped with an NVIDIA RTX 3090 GPU and an Intel i7 CPU. Table 14 presents the breakdown of the computational cost for both 2D-3D and 3D-2D query pipelines. The analysis considers the total time required to process a query frame, including all preprocessing steps (segmentation, depth estimation, projection, inpainting and feature extraction).

In the 3D-2D task (using LiDAR as query sensor), the complete pipeline achieves an inference speed of approximately 9.54 Hz (104.85 ms per frame). It is worth noting that the semantic segmentation backbone (MinkUNet34C Choy et al., 2019) accounts for the majority of the latency. Since the inpainting module is omitted in this modality

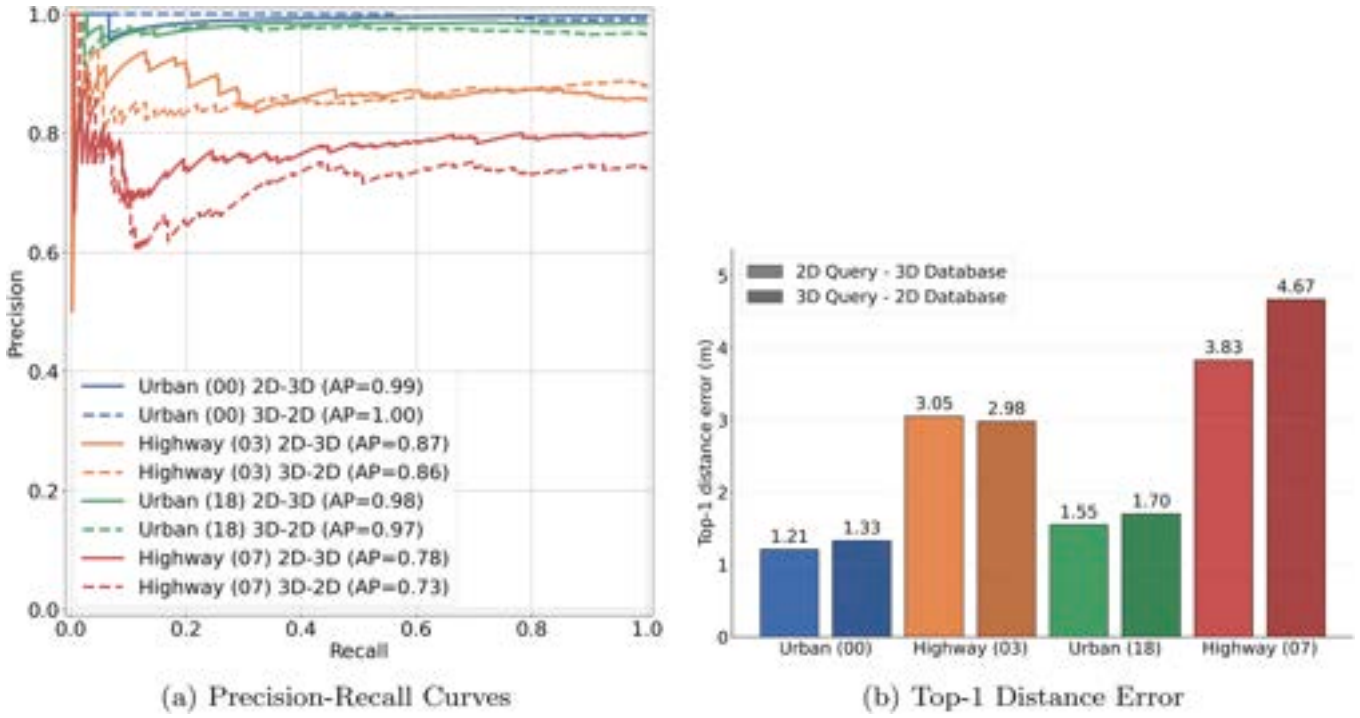

Fig. 9. Performance analysis on KITTI-360 sequences. (a) Precision-Recall curves and (b) The average Top-1 distance error for both 2D-3D and 3D-2D modalities.

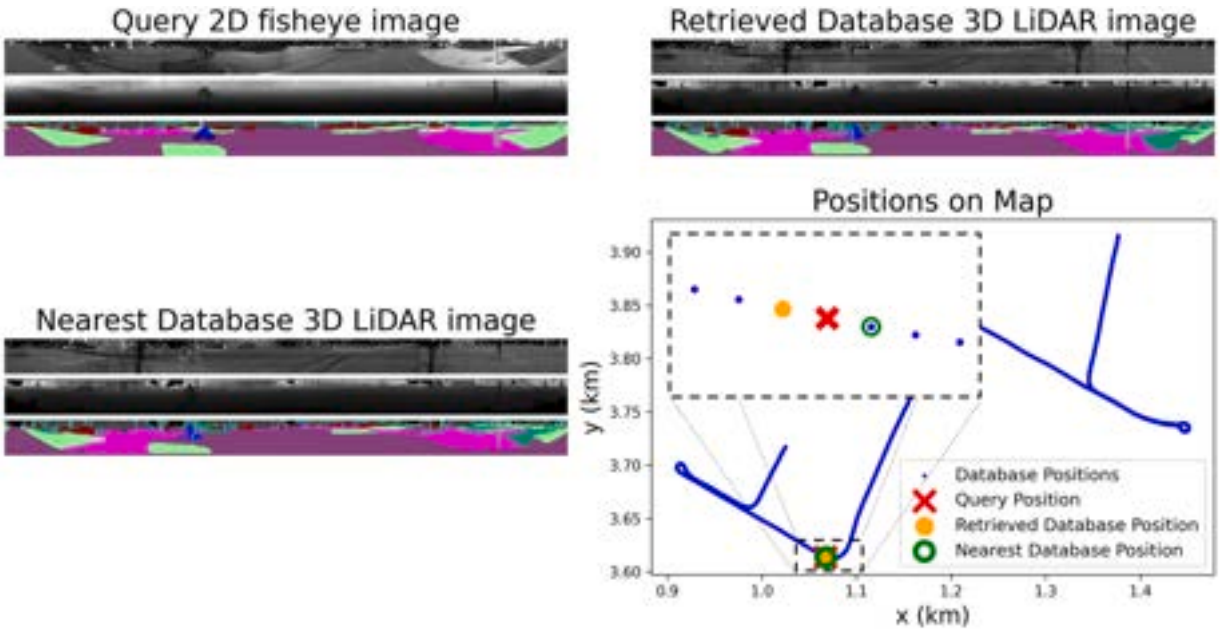

Fig. 10. Example of a correct prediction in 2D-3D modality in environment *00* (urban) with the CrossPlace method in the place recognition task between fisheye images and LiDAR.

(as discussed in Section 4.5.2), the system avoids unnecessary computational overhead, maintaining a frame rate suitable for real-time operation.

Conversely, the 2D-3D (using the camera as query sensor) pipeline operates at approximately 2.19 Hz (456.30 ms per frame). This latency is primarily driven by the use of large-scale, high-accuracy models for semantic segmentation (SegFormer Xie et al., 2021) and depth estimation (Depth Anything V2 Large Yang et al., 2025), which together consume around 406 ms. While these heavy backbones were selected to maximize retrieval performance and establish a robust baseline, they may constitute a system's bottleneck. However, the modular nature of our approach allows these components to be replaced by lightweight alternatives (e.g., SeaFormer++ Wan et al., 2025 or Depth Anything V2 Small Yang et al., 2025) for time-constrained applications, offering a trade-off between accuracy and speed. Additionally, inpainting the ego-vehicle introduces an additional computational cost of approximately 35 ms per frame. For strict real-time robotic applications operating under severe hardware constraints, replacing the inpainting module with a zero-cost static mask stands as a highly viable alternative that preserves competitive retrieval performance while saving critical computational resources. Furthermore, for applications with strict low-latency constraints, the modularity of CrossPlace permits using only the

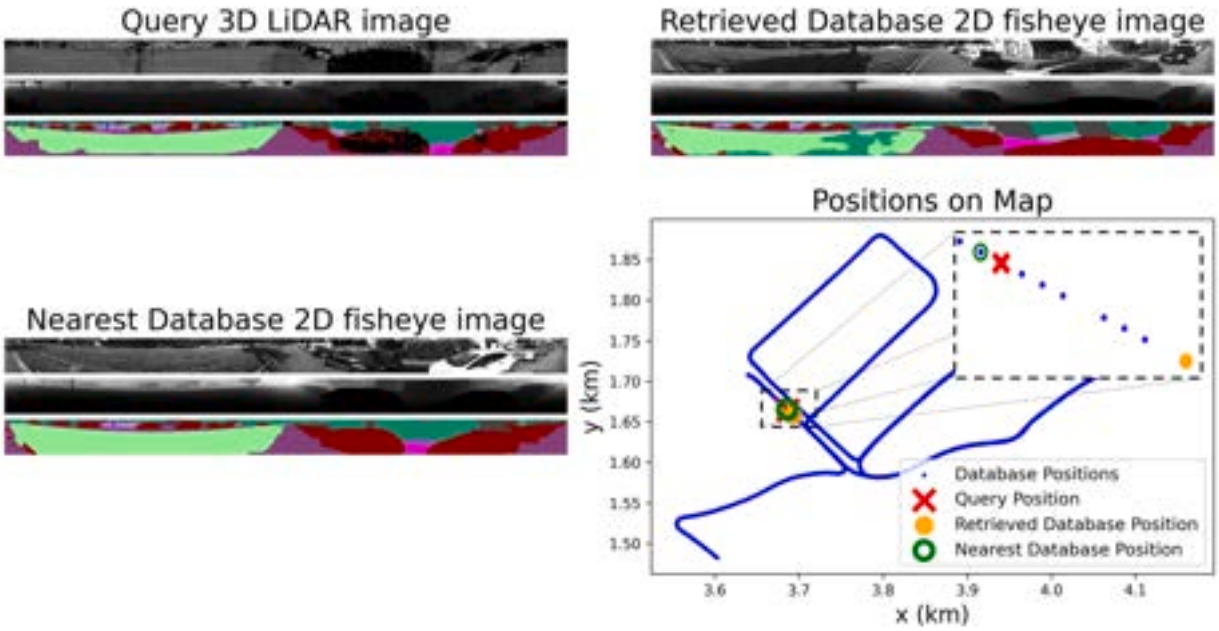

Fig. 11. Example of a slight error in 3D-2D modality in environment *18* (urban) with the CrossPlace method in the place recognition task between fisheye images and LiDAR.

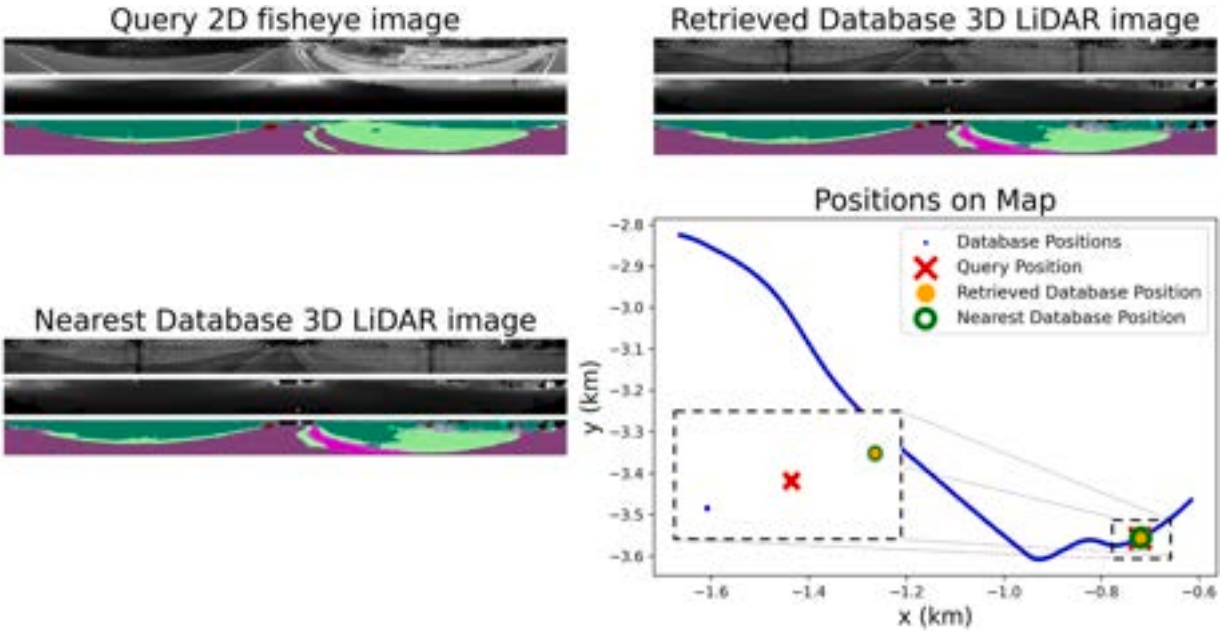

Fig. 12. Example of a correct prediction in 2D-3D modality in environment *07* (highway) with the CrossPlace method in the place recognition task between fisheye images and LiDAR.

intensity branch, which removes the heavy preprocessing steps and enables operation at frequencies higher than 30 Hz for both modalities.

4.7.3. Qualitative results

This section presents visual examples of the performance obtained by the CrossPlace framework proposed in this paper. These examples illustrate the method's capability in urban environments (Figs. 10 and 11) and highway environments (Figs. 12 and 13). In each figure, an example captured in one of the different sequences of the dataset is shown, where the intensity, depth and semantic images from LiDAR or fisheye cameras are shown, along with the prediction of the closest intensity, depth and semantic image in the descriptor space of the database, which is formed by the opposite sensor to the one used for testing. Additionally, it is verified whether it matches with the closest database position, in the metric space. Map positions are represented with blue dots, the current position with a red cross, the predicted position with a yellow circle and the real position with a green ring.

The examples in Figs. 10 and 11 show the method's performance in the urban environment given by sequences *00* and *18*. In Fig. 10, it is observed that the method achieves a correct prediction in the 2D-3D modality, where the semantic segmentations obtained from LiDAR and cameras differ relatively. In Fig. 11, a slight error in the prediction is shown in the 3D-2D modality, where the LiDAR point cloud is slightly shifted with respect to the equirectangular image.

In Figs. 12 and 13, examples from the highway environment are presented, given by sequences *03* and *07*. In Fig. 12, it is observed that the

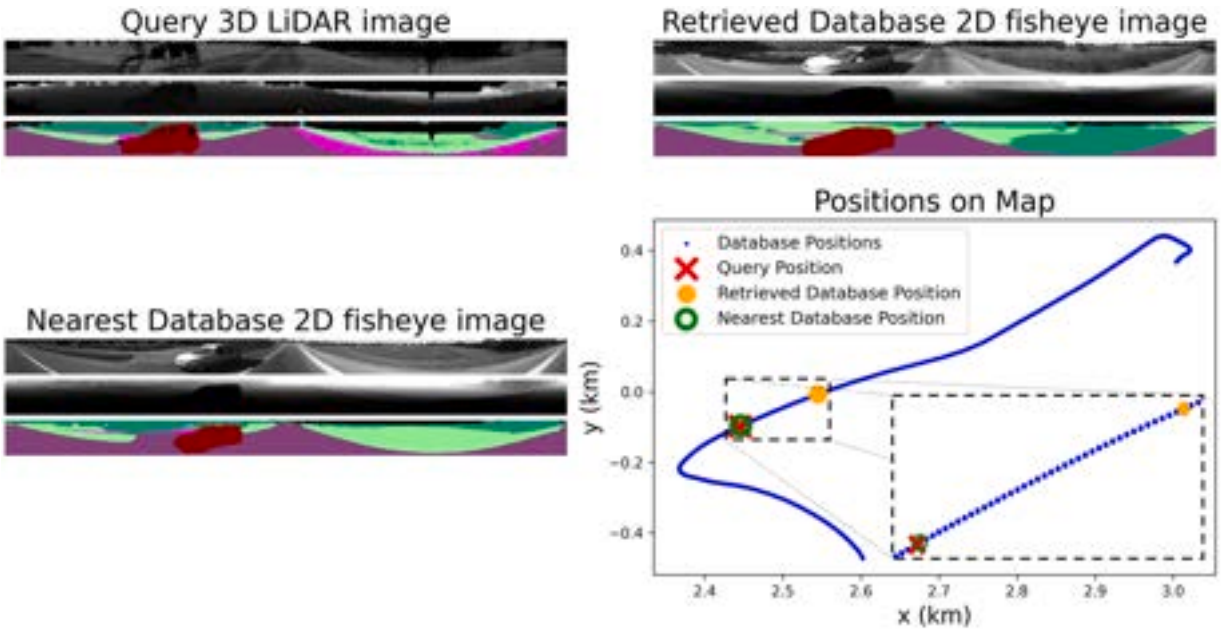

Fig. 13. Example of an error in 3D-2D modality in environment *03* (highway) with the CrossPlace method in the place recognition task between fisheye images and LiDAR.

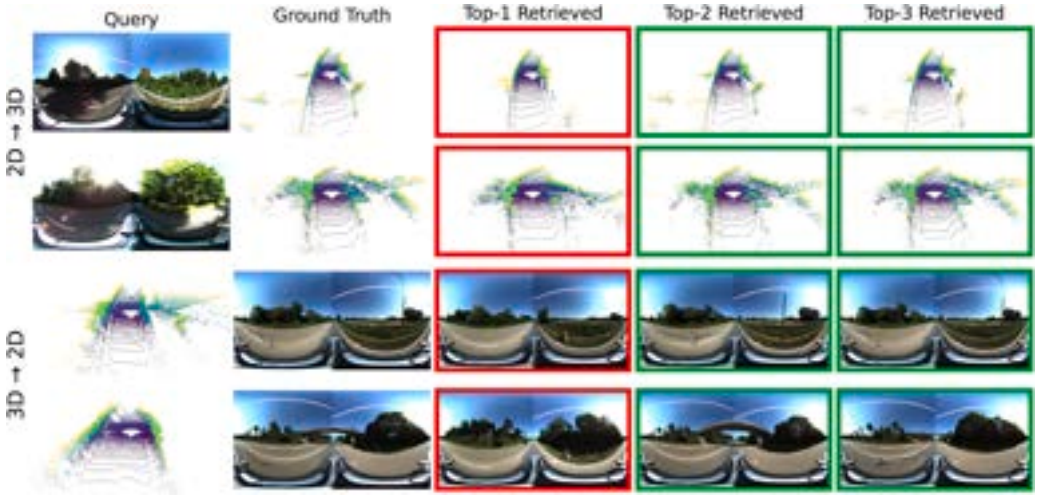

Fig. 14. Examples of failure cases on the KITTI-360 dataset..

method achieves a correct prediction in the 2D-3D modality, even when the semantic segmentation obtained from LiDAR and cameras differs in the semantic segmentation of the sidewalk. In Fig. 13, an example of a failure case in the 3D-2D modality is shown. In this instance, the presence of a vehicle in the semantic image appears to mislead the retrieval process, as the method associates it with another scene containing a vehicle, although it is not the same car. Moreover, in such repetitive environments, it becomes challenging for the method to correctly match test instances with those from the map, since many map instances are highly similar to each other. A wider variety of examples is included on the project website[2].

4.7.4. Limitations

Finally, an exhaustive analysis of the limitations of CrossPlace is performed. Figs. 14 and 15 illustrate several failure cases for the KITTI-360 and NCLT datasets, respectively. Correct predictions are displayed with a green border and incorrect predictions have a red border.

On the KITTI-360 dataset, CrossPlace demonstrates a remarkable performance in urban environments ($R@1$ close to 100%), but shows a slight decrease in highway scenarios. This is primarily attributed to a more pronounced perceptual aliasing, since the scene has the same structure throughout most part of the trajectory. Generally, CrossPlace focuses on discriminative elements, such as traffic signals or changes in the vegetation, that enable a correct retrieval. However, there are some cases in which discriminative elements on the image are not correctly captured by the LiDAR due to its limited vertical FOV, like bridges or overhead power lines (see 3D-2D examples of Fig. 14). It must be noted that, for this dataset, each image and its corresponding point cloud were obtained at the same instant, so the dynamic elements (e.g. cars, pedestrians) appear in both sensor captures. This may lead to cases in which the model benefits from these dynamic elements, which will not be present in real operating situations.

Regarding the NCLT dataset, CrossPlace achieves lower $R@1$ values compared to KITTI-360. For a distance threshold of $d = 5m$, the $R@1$

[2] https://juanjo-cabrera.github.io/projects-CrossPlace/

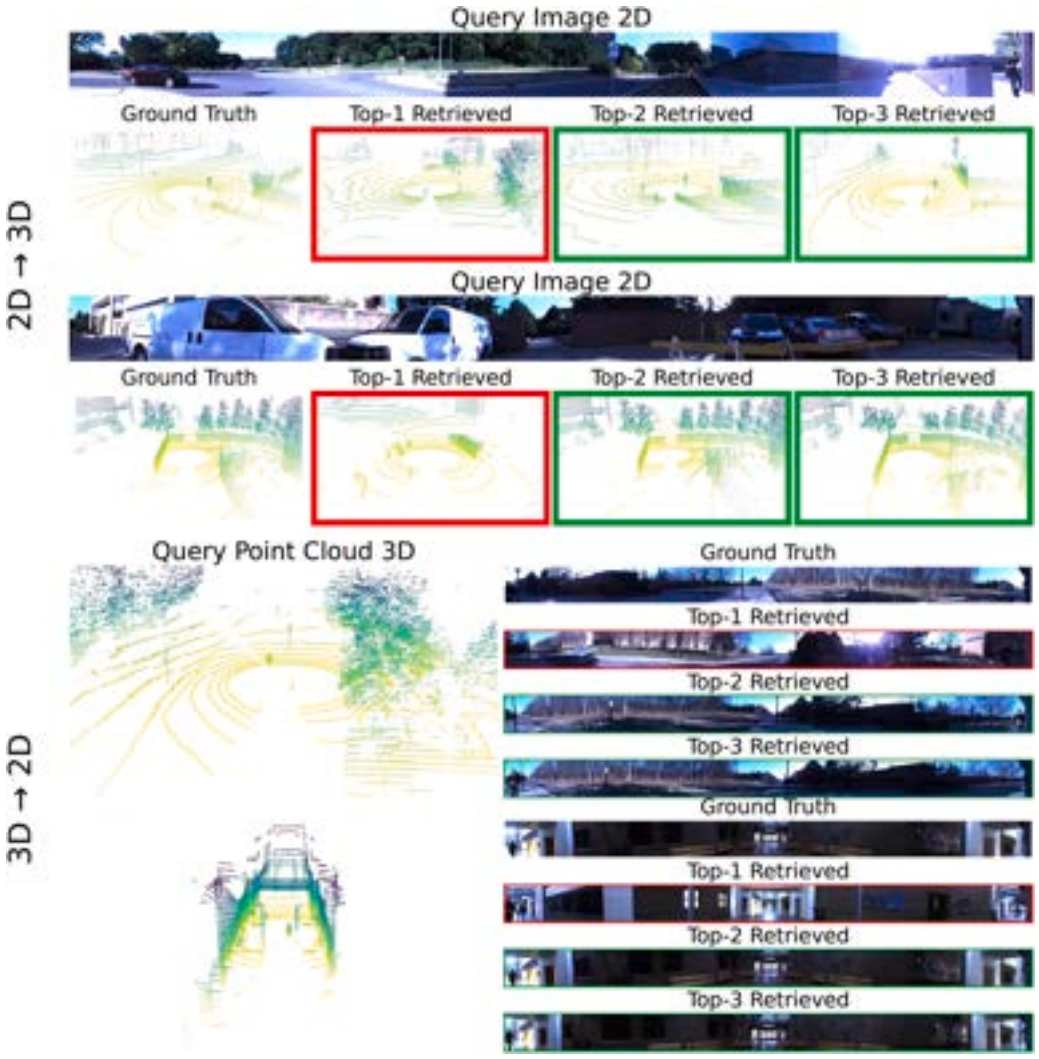

Fig. 15. Examples of failure cases on the NCLT dataset.

is 90.93% on KITTI-360 and 79.57% on NCLT for the 2D-3D modality, whereas for the 3D-2D modality, the $R@1$ is 89.31% on KITTI-360 and 77.55% on NCLT. This underperformance can be explained by several reasons. First, the LiDAR sensor used to build this dataset has a significantly lower resolution, and therefore captures point clouds with lower quality. Second, the omnidirectional view is constructed from six standard cameras instead of high-resolution fisheye cameras, and some of them have a poor white balance setting. Third, each image and its paired point cloud were captured in different recordings that cover the same trajectory. Consequently, there are dynamic elements that appear on the images but not in the point cloud and vice versa, which may affect the retrieval task negatively (see Fig. 15, 2D-3D, second example). Besides, other specific situations can be found, such as translucent surfaces that are not captured by the LiDAR (see Fig. 15, 2D-3D, first example).

Although CrossPlace is not able to retrieve the correct location of the robot in certain situations, most of them are not caused by the model's design nor its training, but by the quality of the sensors or the inherent difficulty of the target environments instead. Besides, it must be noted again that CrossPlace surpasses the state of the art methods on both the KITTI-360 (+51.02% $R@1$ in the 2D-3D modality and +56.61% in the 3D-2D modality with respect to SaliencyI2PLoc Li et al., 2025a) and NCLT datasets (+28.79% $R@1$ in the 2D-3D modality and +22.02% in the 3D-2D modality with respect to InsCMPR Jiao et al., 2025), which are substantially less robust to these situations.

5. Conclusions

This paper has presented an innovative method for cross-modal place recognition between heterogeneous sensor modalities, specifically between fisheye omnidirectional cameras and LiDAR. The proposed method, called CrossPlace, transforms the readings from both sensors into a common space of intensity, depth and semantics, allowing the use of a common three-branch network architecture for both modalities. This approach to place recognition eliminates the need to use the same type of sensor for both queries and database, resulting in a more flexible and practical solution for multi-robot systems or platforms with diverse sensor configurations.

The experimental results on the KITTI-360 dataset have demonstrated the effectiveness of CrossPlace, surpassing all other state-of-the-art methods in all metrics in urban and highway scenarios. In particular, the integration of depth and semantic information has shown to be key for improving discrimination in homogeneous and repetitive environments, such as highways, while intensity information has proven to be an efficient solution without dependence on the estimation of prior learning models.

Furthermore, the impact of different preprocessing techniques, such as vertical interpolation and inpainting, as well as early and late fusion strategies has been evaluated. The results indicate that late fusion through concatenation of intensity, depth and semantic embeddings

provides the best overall performance, highlighting the importance of combining multiple information sources for cross-modal place recognition.

In conclusion, this study introduces a robust and efficient solution for place recognition between different sensor modalities, opening new possibilities for applications in mobile robotics and autonomous systems. As future work, we propose to explore the integration of other sensor modalities, such as thermal sensors and radar sensors, as well as extending the approach to more complex and dynamic scenarios.

CRediT authorship contribution statement

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

Data availability

Data is publicly available on the project website: https://juanjo-cabrera.github.io/projects-CrossPlace/.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Juan Jose Cabrera reports financial support was provided by Spain Ministry of Science and Innovation. Marcos Alfaro reports financial support was provided by Spain Ministry of Science and Innovation. If there are other authors, they 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 FPU21/04969 (J.J. Cabrera) and FPU23/00587 (M. Alfaro). 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).

References

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

Berton, G., Masone, C., & Caputo, B. (2022). Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4878–4888). https://doi.org/10.48550/arXiv.2204.02287

Cai, X., Wang, Y., Huang, Z., Shao, Y., & Li, D. (2024). VOLoc: Visual place recognition by querying compressed LiDAR map. In *2024 IEEE International conference on robotics and automation (ICRA)* (pp. 10192–10199). IEEE.

Carlevaris-Bianco, N., Ushani, A. K., & Eustice, R. M. (2016). University of Michigan North Campus long-term vision and LiDAR dataset. *The* International Journal of Robotics Research, *35*(9), 1023–1035.

Cattaneo, D., Vaghi, M., Fontana, S., Ballardini, A. L., & Sorrenti, D. G. (2020). Global visual localization in LiDAR-maps through shared 2D-3D embedding space. In *2020 IEEE International conference on robotics and automation (ICRA)* (pp. 4365–4371). IEEE.

Choy, C., Gwak, J., & Savarese, S. (2019). 4D Spatio-temporal ConvNets: minkowski convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3075–3084).

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. https://doi.org/10.48550/arXiv.2010.11929

Flores, M., Valiente, D., Peidró, A., Reinoso, O., & Payá, L. (2024). Generating a full spherical view by modeling the relation between two fisheye images. *The* Visual Computer, *40*(10), 7107–7132.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on computer vision and pattern recognition* (pp. 3354–3361). IEEE.

Guan, T., Muthuselvam, A., Hoover, M., Wang, X., Liang, J., Sathyamoorthy, A. J., Conover, D., & Manocha, D. (2023). Crossloc3D: Aerial-ground cross-source 3D place recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11335–11344).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Jiao, S., Su, Z., Luo, L., Yu, H., Zhou, Z., Lu, H., & Chen, X. (2025). InsCMPR: Efficient cross-modal place recognition via instance-aware hybrid mamba-transformer. In *2025 IEEE International conference on robotics and automation (ICRA)* (pp. 2212–2218). IEEE.

Jung, M., Jung, S., Gil, H., & Kim, A. (2025). HeliOS: Heterogeneous LiDAR place recognition via overlap-based learning and local spherical transformer. arXiv:2501.18943.

Karypidis, E., Kakogeorgiou, I., Gidaris, S., & Komodakis, N. (2024). DINO-Foresight: Looking into the future with DINO. https://doi.org/10.48550/arXiv.2412.11673

Komorowski, J. (2022). Improving point cloud based place recognition with ranking-based loss and large batch training. In *2022 26th international conference on pattern recognition (ICPR)* (pp. 3699–3705). IEEE.

Komorowski, J., Wysoczańska, M., & Trzcinski, T. (2021). MinkLoc++: LiDAR and monocular image fusion for place recognition. In *2021 International joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Lai, H., Yin, P., & Scherer, S. (2022). Adafusion: Visual-LiDAR fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, *7*(4), 12038–12045.

Lee, A. J., Song, S., Lim, H., Lee, W., & Myung, H. (2023). LC2: LiDAR-camera loop constraints for cross-modal place recognition. *IEEE Robotics and Automation Letters*, *8*(6), 3589–3596.

Li, Y., Li, J., Dong, Z., Wang, Y., & Yang, B. (2025a). SaliencyI2PLoc: Saliency-guided image-point cloud localization using contrastive learning. *Information Fusion*, *118*, 103015. https://doi.org/10.1016/j.inffus.2025.103015

Li, Y.-J., Gladkova, M., Xia, Y., Wang, R., & Cremers, D. (2025b). VXP: Voxel-cross-pixel large-scale camera-liDAR place recognition. In *2025 International conference on 3D vision (3DV)* (pp. 1233–1242). https://doi.org/10.1109/3DV66043.2025.00117

Liao, Y., Xie, J., & Geiger, A. (2022). KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3292–3310.

Liu, X., Jiang, S., Xie, Y., Lin, Y., & Liu, S. (2026). Modality dominance-aware optimization for embodied RGB-infrared perception. arXiv:2601.00598.

Liu, Y., Chen, G., & Knoll, A. (2020). Globally optimal vertical direction estimation in Atlanta World. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(4), 1949–1962.

Lu, Y., Yang, F., Chen, F., & Xie, D. (2020). Pic-net: Point cloud and image collaboration network for large-scale place recognition. 2008.00658.

Maddern, W., Pascoe, G., Linegar, C., & Newman, P. (2017). 1 Year, 1000 km: The Oxford robotcar dataset. *The International Journal of Robotics Research*, *36*(1), 3–15.

Martínez, Á., Santo, A., Ballesta, M., Gil, A., & Payá, L. (2025). A method for the calibration of a LiDAR and fisheye camera system. *Applied Science*, *15*(4), 2044.

Mei, C., & Rives, P. (2007). Single view point omnidirectional camera calibration from planar grids. In *Proceedings 2007 IEEE international conference on robotics and automation* (pp. 3945–3950). IEEE.

Meng, S., Wang, Y., Xu, H., & Chau, L.-P. (2025). Contrastive learning-based place descriptor representation for cross-modality place recognition. *Information Fusion*, *124*, 103351. https://doi.org/10.1016/j.inffus.2025.103351

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 652–660).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMlR.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision* (pp. 2564–2571). IEEE.

Shubodh, S., Omama, M., Zaidi, H., Parihar, U. S., & Krishna, M. (2024). Lip-loc: LiDAR image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 948–957).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., & Lempitsky, V. (2021). Resolution-robust large mask inpainting with Fourier convolutions. https://doi.org/10.48550/arXiv.2109.07161

Wan, Q., Huang, Z., Lu, J., Yu, G., & Zhang, L. (2025). Seaformer++: Squeeze-enhanced axial transformer for mobile visual recognition. *International Journal of Computer Vision*, *133*(6), 3645–3666.

Wang, S., She, R., Kang, Q., Jian, X., Zhao, K., Song, Y., & Tay, W. P. (2024a). DistilVPR: Cross-modal knowledge distillation for visual place recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10377–10385). (*Vol. 38*).

Wang, W., Min, H., Wu, X., Yang, L., Yan, C., Fang, Y., & Zhao, X. (2024b). Lgd: A fast place recognition method based on the fusion of local and global descriptors. *Expert Systems with Applications*, *251*, 123996.

Xia, Y., Li, Z., Li, Y.-J., Shi, L., Cao, H., Henriques, J. F., & Cremers, D. (2024). UniLoc: Towards universal place recognition using any single modality. arXiv:2412.12079.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

Xie, W., Luo, L., Ye, N., Ren, Y., Du, S., Wang, M., Xu, J., Ai, R., Gu, W., & Chen, X. (2024). ModaLink: Unifying modalities for efficient image-to-pointcloud place recognition. In *2024 IEEE/RSJ International conference on intelligent robots and systems (IROS)* (pp. 3326–3333). https://doi.org/10.1109/IROS58592.2024.10801556

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2025). Depth anything v2. *Advances in Neural Information Processing Systems*, *37*, 21875–21911. https://doi.org/10.48550/arXiv.2406.09414

Yin, P., Xu, L., Zhang, J., Choset, H., & Scherer, S. (2021). i3dLoc: Image-to-range cross-domain localization robust to inconsistent environmental conditions. arXiv:2105.12883.

Zhao, Z., Yu, H., Lyu, C., Yang, W., & Scherer, S. (2023). Attention-enhanced cross-modal localization between spherical images and point clouds. *IEEE Sensors Journal*, *23*(19), 23836–23845.

Zheng, S., Li, Y., Yu, Z., Yu, B., Cao, S.-Y., Wang, M., Xu, J., Ai, R., Gu, W., Luo, L. et al. (2023). I2P-Rec: Recognizing images on large-scale point cloud maps through bird's eye view projections. In *2023 IEEE/RSJ International conference on intelligent robots and systems (IROS)* (pp. 1395–1400). IEEE.

Zhou, W., Jia, M., Lin, C., & Wang, G. (2025). A novel place recognition method for large-scale forest scenes. *Expert Systems with Applications*, *270*, 126606.

Zhou, Y., & Tuzel, O. (2018). VoxelNet: End-to-end learning for point cloud based 3D object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4490–4499).