



**HAL**  
open science

## Extended Study of a Multi-Modal Loop Closure Detection Framework for SLAM Applications

Mohammed Chghaf, Sergio Rodríguez Rodríguez Flórez, Abdelhafid El El  
Ouardi

► **To cite this version:**

Mohammed Chghaf, Sergio Rodríguez Rodríguez Flórez, Abdelhafid El El Ouardi. Extended Study of a Multi-Modal Loop Closure Detection Framework for SLAM Applications. *Electronics*, 2025, 14, 10.3390/electronics14030421 . hal-04906804

**HAL Id: hal-04906804**

**<https://hal.science/hal-04906804v1>**

Submitted on 22 Jan 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Article

# Extended Study of a Multi-Modal Loop Closure Detection Framework for SLAM Applications

Mohammed Chghaf \*, Sergio Rodríguez Flórez \* and Abdelhafid El Ouardi 

SATIE Laboratory—Centre National de la Recherche Scientifique (CNRS)-Unité Mixte de Recherche (UMR) 8029, Paris-Saclay University, Avenue des Sciences Batiment 660, 91190 Gif-sur-Yvette, France; abdelhafid.elouardi@universite-paris-saclay.fr

\* Correspondence: mohammed.chghaf@universite-paris-saclay.fr (M.C.);

sergio.rodriguez@universite-paris-saclay.fr (S.R.F.)

**Abstract:** Loop Closure (LC) is a crucial task in Simultaneous Localization and Mapping (SLAM) for Autonomous Ground Vehicles (AGV). It is an active research area because it improves global localization efficiency. The consistency of the global map and the accuracy of the AGV's location in an unknown environment are highly correlated with the efficiency and robustness of Loop Closure Detection (LCD), especially when facing environmental changes or data unavailability. We propose to introduce multimodal complementary data to increase the algorithms' resilience. Various methods using different data sources have been proposed to achieve precise place recognition. However, integrating a multimodal loop-closure fusion process that combines multiple information sources within a SLAM system has been explored less. Additionally, existing multimodal place recognition techniques are often difficult to integrate into existing frameworks. In this paper, we propose a fusion scheme of multiple place recognition methods based on camera and LiDAR data for a robust multimodal LCD. The presented approach uses Similarity-Guided Particle Filtering (SGPF) to identify and verify candidates for loop closure. Based on the ORB-SLAM2 framework, the proposed method uses two perception sensors (camera and LiDAR) under two data representation models for each. Our experiments on both KITTI and a self-collected dataset show that our approach outperforms the state-of-the-art methods in terms of place recognition metrics or localization accuracy metrics. The proposed Multi-Modal Loop Closure (MMLC) framework enhances the robustness and accuracy of AGV's localization by fusing multiple sensor modalities, ensuring consistent performance across diverse environments. Its real-time operation and early loop closure detection enable timely trajectory corrections, reducing navigation errors and supporting cost-effective deployment with adaptable sensor configurations.



Academic Editor: Yilong Zhu

Received: 3 December 2024

Revised: 15 January 2025

Accepted: 18 January 2025

Published: 21 January 2025

**Citation:** Chghaf, M.; Rodríguez Flórez, S.; El Ouardi, A. Extended Study of a Multi-Modal Loop Closure Detection Framework for SLAM Applications. *Electronics* **2025**, *14*, 421. <https://doi.org/10.3390/electronics14030421>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** camera; LiDAR; fusion; localization; loop closure; mapping; multimodal; particle filter; SLAM

## 1. Introduction

Loop Closure (LC) is a crucial part of SLAM used by autonomous ground vehicles (AGV) [1]. In the context of Simultaneous Localization And Mapping (SLAM) for Autonomous Ground Vehicles (AGV), Loop Closure (LC) serves as a fundamental detection process that reduces the accumulation of errors in motion estimates and enables the solution of the full-SLAM problem within Graph-SLAM frameworks. This aspect is essential because LC involves not only place recognition but also determining the vehicle's current

location on the constructed map. These data are required to correct its previous pose during the Pose-Graph Optimization (PGO) step.

Loop Closure Detection (LCD) can be achieved using two primary techniques: methods based on local descriptors and methods based on global descriptors.

Local descriptor-based techniques rely on extracting descriptors around detected features such as corners or lines from selected keyframes to construct a vocabulary that is quick to create and match. For instance, ORB features extracted from image frames were used in Ref. [2] to build vectors that help recognize a revisited place using the Bag-of-Words (DBOW) scheme [3].

On the other hand, global descriptor-based methods bypass the key points detection phase and aim to describe the entire frame. Inspired by the Scan Context (SC) global descriptor [4], Wang et al. [5] proposed a 3D laser point-cloud descriptor that utilizes intensity information for ground points and height information for non-ground points.

Improving loop closure is crucial for building accurate and consistent maps in SLAM. One effective approach is to use multiple sources of information and multimodal data. For instance, combining data from a camera and LiDAR enhances the system's ability to detect loop closures, as each sensor provides different types of information. This multimodal approach not only leverages the complementary strengths of each sensor but also reduces the impact of errors from individual sensors, thereby increasing the overall robustness of the system [6]. Additionally, integrating different methods of data representation (multiple modalities) from the same information source enriches the representation space and makes the system less vulnerable to missing crucial information.

Overall, utilizing multiple sources of information and multimodal data improves the accuracy and reliability of loop closure detection in SLAM.

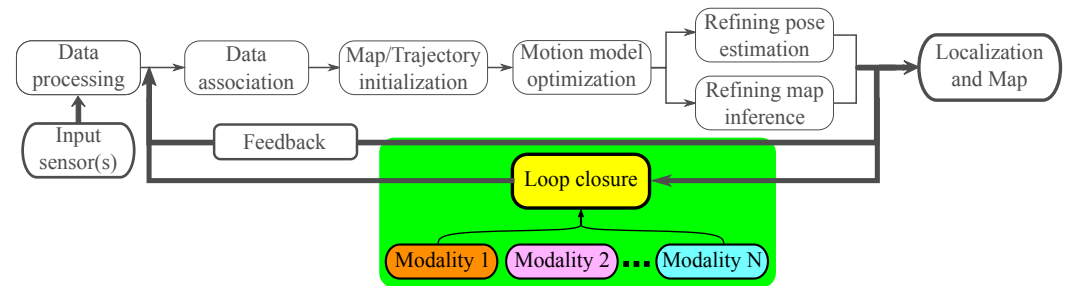
Several challenges in merging multiple modalities for loop closure detection remain unaddressed. Existing fusing schemes presented in [7–10] often tightly couple the representation space of the modalities they employ. This approach can lead to the loss of useful information or incorrect data association. Alternatively, merging multiple modalities in a loosely coupled fashion offers several benefits. Firstly, each modality has an independent role of reliably identifying a loop closure or dismissing false detections from other modalities, especially in challenging and complex situations. Secondly, using the two modalities separately prevents the dominance of one modality over the others. Lastly, accessing information from multiple data streams enhances the system's reliability, redundancy, and fail-safety.

Moreover, although loop closure has achieved great performance thanks to recent advances in the field, to the best of our knowledge, this is the first work aiming at evaluating the impact of accurate and early place recognition on the consistency of the estimated trajectory.

In essence, existing methods often rely on using the same representation space for the employed modalities, making them susceptible to environmental variability such as changes in lighting, weather, or viewpoint. Moreover, many multimodal approaches lack scalability or exhibit high requirements in computational power, limiting their applicability in real-world scenarios like autonomous driving. This study addresses these gaps by proposing a novel Multi-Modal Loop Closure (MMLC) framework that combines complementary descriptors within a Similarity-Guided Particle Filtering (SGPF) mechanism. This fusion strategy enhances loop closure reliability, reduces cumulative localization errors, and ensures robustness across diverse environments and datasets, offering a scalable and efficient solution for real-time AGV applications.

In light of these considerations, this paper presents an extension of our previously published framework [11], which can be integrated into the LC block of any Pose-Graph

Optimization (PGO) based SLAM system, as shown in Figure 1. The proposed methodology leverages multiple modalities to execute MMLC accurately. To demonstrate this concept, we employ our strategy to merge four modalities for LCD purposes.



**Figure 1.** Focus on N different perception modalities into the loop closure module, independent of the used pose-graph optimization-based SLAM system.

In summary, the main contributions of this work are as follows:

1. Characterization of Modalities's Response Model: The response model for each modality used in the approach is characterized, providing a detailed understanding of how different data sources contribute to loop closure detection.
2. Multimodal fusion with an enriched particle filtering strategy: A novel fusion scheme is proposed, based on a similarity-Guided Particle Filter, which infers the most probable loop closure by integrating information from four different modalities.
3. Comprehensive Experiments: Thorough experiments are conducted to validate the approach and compare it to state-of-the-art methods using place recognition and localization metrics on both a publicly available dataset and a self-collected dataset.

The remaining of the article is organized as follows: Section 2 summarizes works related to the subjects at hand. Section 3 outlines the purpose and background of the research while highlighting the choices made and the metrics used for the evaluation. Section 4 briefly describes the proposed and prototyped system. In Section 5 comparative results of the experiments and an in-depth analysis are provided. Finally, Section 6 concludes the article and discusses future perspectives.

## 2. Related Work

According to the literature, the loop closure detection problem has been addressed using a variety of methods. Place recognition in the context of AGV frequently relies on one of two primary sensors: camera or LiDAR. The primary modalities employed in the process to determine if a location has been revisited again rely on visual, geometric/spatial, or semantic data. In the following subsections, we discuss the difference between place recognition and loop closure, then we present state-of-the-art monomodal and multimodal place recognition methods. Then, we present the fusion scheme that will be used in our study.

### 2.1. Place Recognition vs. Loop Closure

Place recognition involves the task of identifying locations that have been visited or noted in the past, forming a crucial component in the sphere of navigation. Achieving accurate recognition requires the use of a database that contains pertinent information about the environments of previously visited places. This information generally comes from various sources including photographs, LiDAR scans, or other sensory mechanisms.

Contrastingly, Loop Closure (LC) is a sophisticated process critical in a SLAM system. It goes beyond identifying a revisited site, which is assured by place recognition techniques. Predominantly, LC is responsible for adjusting the drift in the vehicle's pose estimates as

time progresses. To facilitate this adjustment through pose-graph optimization, a calculation of the relative transformation between the two poses illustrating an LC is required. Notably, these kinds of geometric data are not supplied by the current general-purpose algorithms implemented in place recognition, as outlined in [12].

Furthermore, techniques utilized in place recognition predominantly aim at identifying locations that have been revisited across a comprehensive set of frames within a dataset. Conversely, within the framework of SLAM, keyframes are typically employed to pinpoint a closed loop. A statistical analysis depicting the comparison between revisited frames and closed loops, utilizing the KITTI dataset, is presented in Table 1 [13]. The methodology for keyframe detection adheres to the procedures delineated in [2].

**Table 1.** Comparison of revisited locations and loop closure detection metrics with ORB-SLAM2 [2] and KITTI dataset [13].

Sequence	Total Frames	Revisited Frames	Total Keyframes	Revisited Keyframes	Invalid Relative Transformation	Cancelled LC	Effective LC
KITTI_00	4541	776	1392	151	79	4	5
KITTI_02	4661	299	1748	83	45	26	3
KITTI_05	2761	425	570	58	67	0	2
KITTI_06	1101	268	505	94	19	46	3
KITTI_07	1101	28	253	4	2	0	1
KITTI_08	4071	321	1209	105	0	0	1

The aforementioned table enumerates both the number of revisited frames and the count of revisited keyframes. Drawing upon the data procured from [2], the table additionally describes the quantity of identified Loop Closures (LCs) that were disregarded due to the erroneous relative pose transformations, which failed to meet the criteria for a consistency check. The final two columns document the instances of terminated LCs during the phase of pose-graph optimization, along with those that reached successful completion.

This analytical representation elucidates the performance capability of a place recognition algorithm such as DBoW2 [3] in facilitating the Loop Closure Detection (LCD) procedure within a SLAM system. Moreover, it accentuates the notion that a precise detection of a revisited location does not unequivocally guarantee a successful loop closure in the domain of SLAM.

The most advanced place identification systems take into consideration the whole set of frames in the dataset and are confined to a binary classification issue. However, in the context of SLAM, the difficulties are not restricted to finding revisited locations from a set of keyframes (which is, by definition, smaller than the set of frames). But also successfully estimating a similarity transformation using the detected LC keyframe.

## 2.2. Place Recognition

### 2.2.1. Place Recognition Using Camera

Modern cameras offer vast potential for generating high-quality and abundant data suitable for loop closure detection. By leveraging this sensor, various strategies have been developed for place recognition, broadly categorized as image-to-image, image-to-map, and map-to-map methods [14]. Image-to-image methods are commonly applied in SLAM systems, often employing the Bag-of-Words (BoW) model to simplify computational demands. The BoW approach converts an image into a set of descriptors, from which a BoW vector is generated and used to match images against previously registered vectors. Galvèz et al. [3] introduced a method utilizing ORB features to construct a vocabulary of binary words, enabling the faster creation and matching of descriptors. Building on this, ORB-SLAM2 [2] extracts ORB features from selected keyframes to generate vectors, which

are then used to identify loop closure candidates based on the BoW scheme [3]. To validate the loop, this approach incorporates geometric verification by calculating the similarity transformation between the current keyframe and the identified candidate.

In Ref. [15], NetVLAD is introduced. It is based on an end-to-end Convolution Neural Network (CNN) and is able to transform an image into a Vector of Locally Aggregated Descriptors (VLAD). However, this method does not take into account dynamic objects in the environment when creating a global descriptor, which can lead to mismatches.

To overcome this problem, Zhang et al. in [16] presented a novel loop closure detection approach based on image inpainting and feature selection. Thus, only valid superpoint features in the areas with high inpainting qualities are selected as the input of the Bag of Words model for loop-closure detection.

More recent results, such as the work of Xiao et al. [17], proposed efficiently recognizing loop closures by leveraging the semantic information contained in monocular images. This approach can be generalized to panoramic images while preserving the panorama's characteristics, which outperforms monocular detector-descriptor-based algorithms. Although this method can significantly increase the accuracy of loop closure detection tasks, it is not suitable for real-life SLAM applications. Indeed, it is time-consuming and does not provide relative pose estimation to correct the drift in the trajectory.

### 2.2.2. Place Recognition Using LiDAR

Camera-based place recognition is highly sensitive to appearance changes, making it vulnerable to errors under significant viewpoint variations or environmental changes. In contrast, LiDAR, while limited to providing geometric information and intensity measurements, offers advantages such as 360° 3D scans and extended range compared to cameras. These features make LiDAR an excellent alternative for detecting loop closures in AGV applications, effectively addressing the limitations of camera-based approaches.

Recent advancements in Convolutional Neural Networks (CNNs) have enabled innovative methods like PointNetVLAD [18], which tackles large-scale place recognition by extracting local features and clustering them into a VLAD global descriptor through the integration of NetVLAD and PointNet.

Another promising method, Scan Context [4], introduces a spatial global descriptor that encodes 3D point clouds into a matrix. This matrix captures the structural details of the scene by leveraging the height, azimuthal, and radial information of the points, demonstrating strong potential for place recognition tasks. This global descriptor used for loop closure detection is especially used in recent lightweight SLAM systems such as E-LOAM [19] and help efficiently correct the trajectory drift.

Recently, Xiang et al. [20] proposed a very deep and lightweight Siamese feature extraction module and a dual-attention-based feature difference module that can perform real-time and reliable LCD in large-scale environments.

### 2.3. Multimodal Loop Closure Detection

Taking advantage of a unified framework to combine multiple sources of information has been shown to significantly enhance SLAM performance, particularly in place recognition tasks [1,6,21]. Early approaches, such as the work by Collier et al. [22], introduced a fusion of features extracted from both camera and LiDAR data using parallel BoW vocabularies. Each sensor independently validates loop closure candidates through a 6-DoF transformation, with the results combined for geometric loop closure. While effective, this method requires a computationally intensive training step, limiting its applicability in real-time scenarios.



More recent methods, such as CORAL [9], adopt a bi-modal representation to integrate visual and structural features into a bird's-eye view. This strategy provides an innovative representation but lacks the flexibility to adapt to highly dynamic environments. MinkLoc++ [8] takes a different approach by aggregating point cloud and RGB image descriptors through a late fusion technique. Although effective in descriptor generation, the network's reliance on the modality with superior training performance (RGB images) results in suboptimal outcomes. The authors address this by incorporating uni-modal descriptors into the loss function, but the method's inability to estimate relative pose limits its integration into SLAM pipelines.

Alternatively, semantic fusion methods like SVG-Loop [23] use a combination of visual-based and semantic-based similarity scores to validate loop closures. While effective in some cases, this technique heavily depends on segmentation accuracy and struggles with significant viewpoint changes, reducing its robustness in diverse settings.

Traditional monomodal loop closure detection methods, such as those based on visual or geometric descriptors, have demonstrated effectiveness under controlled conditions. However, these methods often struggle with environmental variability, like the lack of scene structure or texture, which limits their robustness in diverse real-world scenarios. On the other hand, multimodal approaches aim to overcome these limitations by fusing complementary sensory data. For instance, learning-based methods excel at recognizing high-level features but require extensive training datasets and can be sensitive to domain shifts. Conversely, geometric-based methods are more stable across datasets but may lack the representational richness of learned features. Despite these advances, existing multimodal techniques often focus on sensor fusion without addressing computational efficiency or scalability, particularly for embedded systems in autonomous vehicles. Our work bridges these gaps by proposing a unified framework that leverages the strengths of both monomodal and multimodal approaches, combining learned and geometric descriptors using a robust particle filtering mechanism to improve accuracy and scalability in diverse environments.

### 3. Methodology

In this section, we introduced the purposes of fusing multiple Loop Closure Detections (LCD) in the SLAM context. Then, we justified the choices made in terms of the SLAM algorithm and place recognition methods. Finally, we unveiled the evaluation metrics that were subsequently used in the experimental results to assess the impact of our contributions.

#### 3.1. Research Purpose

The existing place recognition methods are robust enough when they face environments similar to those on which they had been tested and validated. However, their resiliency drops drastically, especially those based on deep learning strategies, when facing completely new challenging environments. Moreover, on one hand, the existing local descriptors are robust enough for rotation, but they suffer from a lack of efficiency in environment description. On the other hand, global descriptors can successfully describe a scene but will lead to mismatches when faced with rotations. Furthermore, tightly coupling the modalities used for place recognition leads to losing information in the process or to errors in data association. This can be avoided by using a loosely coupled strategy where every modality is processed independently. On top of that, such a strategy can overcome scenarios where a source of information is temporarily unavailable or faulty.

Finally, studying the loop closure problem and quantifying the impact of accurately detecting and correcting the loop on the resulting trajectory in the SLAM context can help design systems that are more suited for specific applications, such as AGVs.

Thus, combining descriptors that leverage multiple information/ modalities/ strategies will help increase the redundancy of the SLAM system, while guaranteeing its robustness and the consistency of its estimated trajectory.

### 3.2. SLAM Algorithm Choice

This study's framework was part of the design of an embedded multimodal SLAM system for AGV applications. ORB-SLAM2 [2] was chosen as a baseline since it is an open-source and stable algorithm that provides satisfactory results in outdoor dynamic scenes. Moreover, it can be embedded [24,25]. Thanks to its low complexity and modularity, ORB-SLAM2 will help our study by focusing mainly on the LCD strategy without worrying about the front end of SLAM. Furthermore, ORB-SLAM2 uses ORB features and Bag-of-Words [3] to detect LCs, which was one of the modalities we used in the proposed multimodal LC.

### 3.3. Loop Detection Methods Choice

In light of our research purposes, the chosen place recognition methods to be merged should meet three criteria: First, use camera or LiDAR information as they are the most widely used in SLAM applications for AGVs. Second, we represent the processed information in different spaces. Third, we assess the public availability of the source code used to test and validate such a method. As a Proof-of-Concept (PoC) of the proposed framework, we selected the following place recognition methods:

1. DBOW: Will be referred to in the following DBOW or D. Uses ORB features detected on camera frames.
2. NetVLAD: Will be referred to in the following NV or N. Uses Convolutional Neural Network (CNN) to create a Vector of Locally Aggregated Descriptors (VLAD) that represents a camera frame.
3. Scan Context: Will be referred to in the following SC or S. Uses geometric information contained in the LiDAR point cloud to create a global descriptor of the current frame.
4. PointNetVLAD: Will be referred to in the following PNV or PN. Uses Convolutional Neural Network (CNN) to create a global descriptor from a given 3D LiDAR point cloud.

### 3.4. Evaluation Metrics

Since the proposed method is designed to infer the most reliable Loop Closure Candidate (LCC) in SLAM applications; the chosen evaluation metrics leverage performance indicators used in both place recognition research and in the SLAM context. The following is a list of the metrics we relied on during this study:

1. AUC: measures the Area Under Curve given by the Precision-Recall. It evaluates a model's capability to differentiate between negative and positive classes.
2. F1: the maximum F1 score [26] is the harmonic mean of Precision and Recall. Defined as:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (1)$$

This metric quantifies overall classification performance by treating precision and recall as equally important.

3. EP: the Extended Precision [27] is defined as follows:

$$EP = 0.5 \times (P_{R0} + R_{P100}) \quad (2)$$

Here,  $P_{R0}$  represents the precision at the minimum recall, while  $R_{P100}$  indicates the maximum recall achievable at 100% precision. By combining these two indicators, this



metric provides valuable insight into the method's performance. Specifically,  $R_{P100}$  reflects the highest level of recall attainable without introducing any False Positives (FP), which is crucial as even a single FP can significantly degrade the pose graph during the loop correction phase of a SLAM system.

4. ATE: Absolute Translation Error proposed in [28] estimates the global consistency of the estimated trajectory when using a SLAM system. This evaluation is carried out by comparing absolute distances (in meters) between the estimated trajectory and the ground truth (GT). This error can be computed by taking into account the rigid-body transformation  $S$  that provides the least-squares solution to align the estimated trajectory  $P_{1:n}$  with the ground truth trajectory  $Q_{i:n}$ , it is defined as follows:

$$ATE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left\| \text{trans}(Q_i^{-1} S P_i) \right\|^2} \quad (3)$$

5.  $t_{rel}$  and  $r_{rel}$ : the average *Relative Translation* (in %) and *Relative Rotation* (in  $^\circ/100\text{m}$ ) errors proposed in [13] evaluate the local accuracy of the trajectory over a fixed trajectory length. These error metrics are defined as follows:

$$t_{rel} = \frac{1}{\mathcal{F}} \sum_{(i,j) \in \mathcal{F}} \angle \langle (\hat{p}_j \ominus \hat{p}_i) \ominus (p_j \ominus p_i) \rangle \quad (4)$$

$$r_{rel} = \frac{1}{\mathcal{F}} \sum_{(i,j) \in \mathcal{F}} \left\| (\hat{p}_j \ominus \hat{p}_i) \ominus (p_j \ominus p_i) \right\|_2 \quad (5)$$

Where  $\mathcal{F}$  is a set of frames  $(i, j)$ ,  $\hat{p} \in SE(3)$  and  $p \in SE(3)$  are estimated and GT poses, respectively.  $SE(3)$  is the Special Euclidean Group that represents rigid body transformations in 3D space, including both rotation and translation.  $\angle \langle \cdot \rangle$  is the rotation angle and  $\ominus$  is the inverse motion compositional operator [29] that enables the computation of the relative transformation that moves the node given by pose  $p_i$  to  $p_j$ .

From a system level, the key aspects that need to be assessed in multimodal SLAM are its ability to correctly close the loop when a place is revisited, and its capability of accurately estimating the trajectory and reducing the cumulative errors during runtime. From one perspective, metrics such as AUC, F1 and EP have been commonly used in the literature for the rare events of revisiting a place where the classification problem is skewed [30]. These metrics measure the robustness of the LCD and capture the system's ability to distinguish a small set of revisited places from a larger set of initial visit locations. From another perspective, metrics like ATE,  $t_{rel}$  and  $r_{rel}$  focus on trajectory accuracy by quantifying the deviation between estimated and ground-truth trajectories, highlighting the system's ability to preserve global and local consistency [28]. Unfortunately, these trajectory metrics focus on the SLAM system performance at the end of the trajectory. The AGV application constraints require the error to be minimized as quickly as possible. Therefore, we introduced a metric that evaluates the system's ability to reduce the error both efficiently and rapidly during the execution of the algorithm, rather than solely at its completion. This metric is critical in SLAM because it directly affects the accuracy and consistency of the generated map and the system's localization capabilities. As SLAM systems operate in real-time, errors in trajectory estimation or place recognition can compound over time, leading to significant deviations from the true path and reducing the system's chance to close loops effectively.

## 4. System Description

In this section, we introduce the Similarity-Guided sampling mechanism, which employs a Gaussian-mixture proposal to detect loop closures using various perception modalities. The modes of the posterior Probability Density Function (PDF) are estimated by evaluating the similarity between the current keyframe and the entire set of keyframes.

### 4.1. Problem Formalization

Considering the set of all prior keyframes during an AGV's trajectory, denoted as  $\mathcal{F}_t = \{F_i, |, i = 0, \dots, t\}$ , we define  $\overset{*}{\mathcal{F}}_t = \{F_i, |, i = 0, \dots, t - k\}$  as the subset of keyframes where a loop closure can potentially occur. This assumes the AGV begins at time step 0 and is currently at time step  $t$ , with  $\llbracket t - k, t \rrbracket$  representing a time window where loop closures are not possible. Here,  $k$  is a fixed small number of recently visited keyframes that cannot correspond to a revisited location.

Let  $\mathcal{L}^i = \{L_1^i, L_2^i, \dots, L_{m_i}^i\}$  represent the set of loop closure candidates identified by perception modality  $i$ , where  $i \in \llbracket 1, N \rrbracket$ , assuming  $N$  modalities are used for loop closure detection. The corresponding normalized similarity scores are given by  $\Sigma^i = \{\sigma_1^i, \sigma_2^i, \dots, \sigma_{m_i}^i\}$ . Notably,  $\mathcal{L}^i \subset \overset{*}{\mathcal{F}}_t$ .

Using these definitions, our goal is to estimate the unknown variable  $\mathcal{X}$ , which represents all loop closure events throughout the vehicle's trajectory. This variable depends on the similarities with previously observed poses. The problem can thus be formulated as follows:

$$\chi^* = \underset{\chi}{\operatorname{argmax}} \mathcal{P}(\chi|\Sigma) = \underset{\chi}{\operatorname{argmax}} \mathcal{P}(\Sigma|\chi)\mathcal{P}(\chi) \quad (6)$$

The equality follows from the Bayes theorem in Equation (6) where  $\mathcal{P}(\Sigma|\chi)$  is the likelihood of the similarities  $\Sigma$  given the assignment  $\chi$  and  $\mathcal{P}(\chi)$  is a prior probability over  $\chi$ . Assuming that these similarities are independent (the corresponding noises are uncorrelated), Equation (6) factorizes into

$$\chi^* = \underset{\chi}{\operatorname{argmax}} \mathcal{P}(\chi) \prod_{i=1}^M \mathcal{P}(\sigma_i|\chi) \quad (7)$$

### 4.2. Similarity-Guided Particle Filtering (SGPF)

To address Equation (7), we propose leveraging a Particle Filter (PF) to identify the keyframe with sufficient evidence to confirm a loop closure event. Building upon previous works accomplished in [31,32], we resample particles based on the appearance similarity between the current keyframe and keyframes stored in the map.

Considering  $M$  similarity-based particles  $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_M\}$  derived from  $N$  perception modalities. The probability of a loop closure occurrence at the time step  $t$  can then be expressed as follows:

$$\mathcal{P}(x_t|\Sigma_t) = \mathcal{P}(x_t) \prod_{i=1}^M \mathcal{P}(\sigma_i|x_t) \quad (8)$$

Where the set of all similarities fixed by the  $N$  perception modalities is given by the following:

$$\Sigma_t = \bigcup_{i=1}^N \Sigma_t^i \quad (9)$$

To solve Equation (8), the posterior distribution  $\mathcal{P}(x_t|\Sigma_t)$  is represented using a set of weighted loop closure candidates  $\mathcal{L}t = \{(L_t^i, \omega_t^i), |, i = 1, \dots, M\}$  at each time step  $t$ . Here,  $L_t^i$  denotes a particle representing a proposed loop closure keyframe, and  $\omega_t^i$  is its associated weight. These weights are determined using an importance sampling approach [33], where  $\omega_t$  is defined as  $\frac{\mathcal{P}(x_t|\Sigma_t)}{\mathcal{Q}(x_t|\Sigma_t)}$ . The term  $\mathcal{Q}(x_t|\Sigma_t)$  represents a proposal function, designed to facilitate easier sampling compared to directly sampling from the posterior density  $\mathcal{P}(x_t|\Sigma_t)$ . In this framework, the proposal function is expressed as  $\mathcal{Q}(x_t|\Sigma_t) = \mathcal{P}(x_t|\mathcal{X}_{0:t-1}, \Sigma_t)$ , incorporating the importance of  $\Sigma_t$  in the sampling process. Where  $\mathcal{X}_{0:t-1}$  is the set of LC occurrences in the vehicle trajectory up to current frame given by index  $t - 1$ . The weights are proportional to the similarity score and the loop closure occurrence likelihood. For each particle  $L_t$ , it can also be written as follows:

$$\omega_t \propto \sigma_t \times \mathcal{P}(\sigma_t|x_t) \quad (10)$$

With the theoretical foundation of the Similarity-Guided Particle Filter (SGPF) established, we now outline the process for generating particles. Initially, a set of particles representing LCDs from  $N$  perception modalities is generated, denoted as  $\mathcal{L}t = \{\mathcal{L}t^k, |, k = 1, \dots, N\}$ . This set contains  $|\mathcal{L}t| = M_1$  candidates produced by the perception modalities. These candidates are sampled using a Gaussian distribution with covariance matrix  $\mathcal{G}1$ :  $\mathcal{P}(x_t|\sigma_i) \sim \mathcal{N}(\sigma_i, \mathcal{G}_1)$ .

To further enrich the pool of candidates, an additional set is generated, represented as  $\hat{\mathcal{L}}t = \{\hat{\mathcal{L}}t^k, |, k = 1, \dots, N\}$ . This set comprises  $|\hat{\mathcal{L}}t| = M_2$  candidates sampled from Gaussian distributions with covariance matrix  $\mathcal{G}2$ :  $\mathcal{P}(x_t|\sigma_i) \sim \mathcal{N}(\sigma_i, \mathcal{G}_2)$ . Together, these sets provide a comprehensive and diverse pool of loop closure candidates.

The mixture distribution outlined in Equation (11) is ensured by defining the total number of particles as  $M = M_1 + M_2$ . Consequently, at time step  $t$ , the complete sample set is represented as  $\hat{\mathcal{L}}t^* = \mathcal{L}t \cup \hat{\mathcal{L}}t$ .

$$\mathcal{P}(x_t|\Sigma_t) = \sum_{i=1}^M \mathcal{P}(x_t|\sigma_i) \quad (11)$$

Consequently, the final set of particles for loop closure detection consists of a combination of similarity-based candidates and model-generated particles sampled around these proposed candidates. This entire workflow is outlined in Algorithm 1. The SGPF framework begins by initializing with the  $t - k + 1$  keyframes from the map, potential LCD identified by place recognition methods, and additional model-based particles sampled around these initial candidates. Following this, the similarity scores for all particles are updated and normalized, and each particle's weight is computed based on the confidence associated with its respective modality. Subsequently, during the resampling phase, particles with lower weights are eliminated, while those with higher weights are replicated, ensuring the particle filter remains robust and avoids degeneracy. Finally, the loop closure process and pose-graph optimization are executed only when the weight of a selected particle surpasses a predefined threshold.

To clarify the methodology, Figure 2 provides a step-by-step illustration of the particle filtering process employed in the SGPF approach. The figure outlines the key stages of the process, starting from candidate generation, followed by similarity computation and particle propagation, to the final validation of LCD and the resampling step.

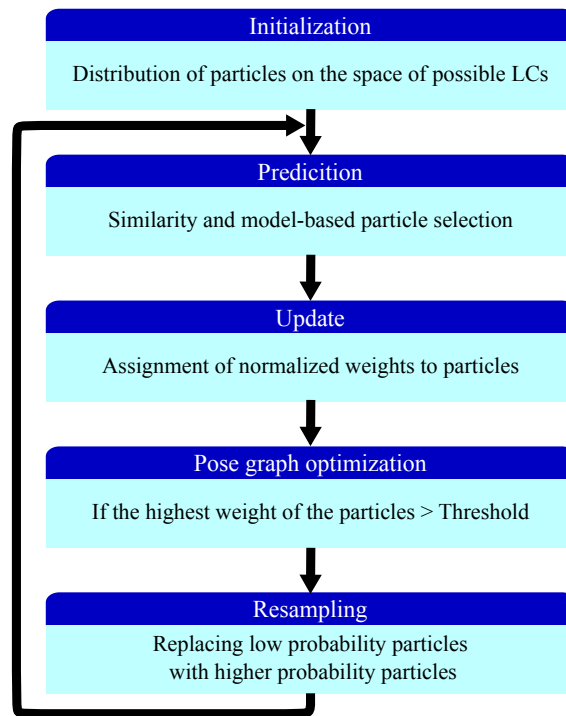


Figure 2. Step-by-step diagram of the SGPF.

---

**Algorithm 1:** Loop closure detection pseudo-code using Similarity-Guided Particle Filter

---

**Input** :  $t - k + 1$  : Keyframes in the map  
 $M_1$  : proposed particles  
 $M_2$  : model-based particles

**Output** : Predicted loop closure keyframe index

```

1 /* Step 1: Initialization */
2 Generate initial particles according to Equation (8)
3 begin
4   /* Step 2: Particles propagation */
5   Generate  $\mathcal{L}_t$  candidates based on their similarity score;
6   Predict  $\hat{\mathcal{L}}_t$  particle based on the proposed particles;
7    $\hat{\mathcal{L}}_t^* = \mathcal{L}_t \cup \hat{\mathcal{L}}_t$ ;
8   /* Step 3: Scores update and normalization */
9   foreach  $L_t^i$  in  $\hat{\mathcal{L}}_t^*$  do
10    |  $s_t^i = \text{Similarity}(L_t^i, \text{CurrentKeyframe})$ ;
11    |  $\hat{s}_t^i = \text{Normalize}(s_t^i)$ ;
12    |  $\text{ScoresSum} = \text{ScoresSum} + \hat{s}_t^i$ ;
13  end
14  /* Step 4: Weights normalization */
15  foreach  $L_t^i$  in  $\hat{\mathcal{L}}_t^*$  do
16  |  $\omega_t^i = \hat{s}_t^i / \text{ScoresSum}$ ;
17  end
18  /* Step 5: Resampling */
19  Resample( $\hat{\mathcal{L}}_t$ );
20   $x_t^* = \underset{\chi}{\text{argmax}} \mathcal{P}(x_t | \omega_t)$ ;
21  if  $\omega_t^* > \text{threshold}$  then
22  | PoseGraphOptimization();
23  end
24 end
  
```

---

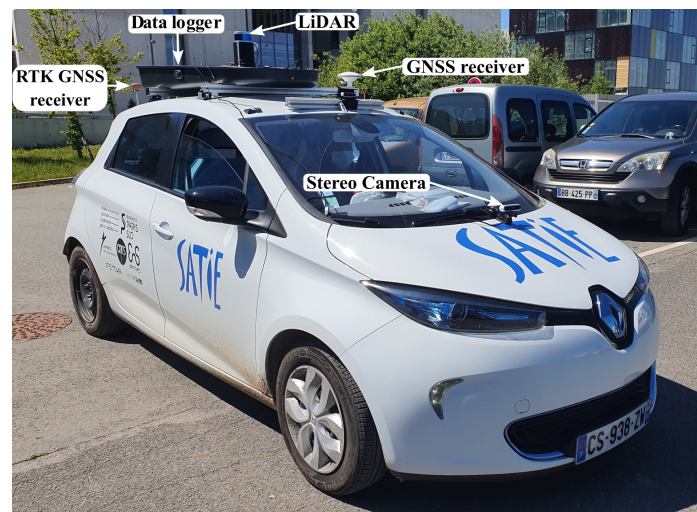
## 5. Experimental Results and Analysis

In this section, the proposed method is implemented by using four loop detection modalities: Bag-of-Words (DBOW) [3], NetVLAD (NV) [15], Scan Context (SC) [4] and PointNetVLAD (PNV) [18]. First, we describe the datasets and experimental settings used to validate our approach. Then, we illustrate the correlation between the ground truth distance and modality distance between a current keyframe and a Loop Closure Candidate (LCC). Finally, we perform qualitative and quantitative extended evaluations of our approach.

### 5.1. Datasets Description and Experimental Setup

The proposed framework was tested and validated on all the sequences from the KITTI dataset [13] that include loop closures. Those sequences are 00, 02, 05, 06, 07, 08. As described in [34], the LiDAR and camera's temporal synchronization is hardwired, with the cameras being triggered when the LiDAR scanner is oriented toward the scene in front of the car. Both camera frames and LiDAR frames are acquired at 10 Hz.

In order to further investigate the modalities complementarity to avoid falling into the trap of modality one dominance, the instrumented car from the SATIE laboratory (shown in Figure 3) was used to collect data from three sequences containing loops (total distance, respectively, 2 km, 1.08 km and 0.75 km). The vehicle has a camera (Intel Realsense D455), a LiDAR (Velodyne VLP-16), and a GNSS receiver (Altus Positioning System) that uses Real-Time Kinematic (RTK) corrections to offer precision at the centimeter scale. While the LiDAR data were recorded at 10 Hz, the stereo frames were obtained at a rate of 30 Hz.



**Figure 3.** SATIE Laboratory instrumented car. The vehicle is equipped with a stereo-camera, a LiDAR and a data logger. The ground truth trajectory is recorded using an RTK GNSS receiver.

Data were recorded using ROS (Robot Operating System) and sensor calibration was conducted prior to data collection using MATLAB Version R2020b. To achieve sensor synchronization, for each LiDAR frame, the corresponding camera frame with the closest timestamp was selected. The data collection process involved driving the vehicle through urban and suburban environments under varying traffic and weather conditions to ensure diversity. The vehicle's trajectories were chosen to include both dynamic (e.g., cars, buses, pedestrians) and static elements (e.g., buildings, houses, trees). Figure 4 illustrates a selection of the trajectories collected during the process. Additionally, several frames were largely dominated by the sky or the shade. All experimental validation and tests were conducted on a workstation with an Intel i7-11800H CPU (2.30 GHz) and 64 GB memory.

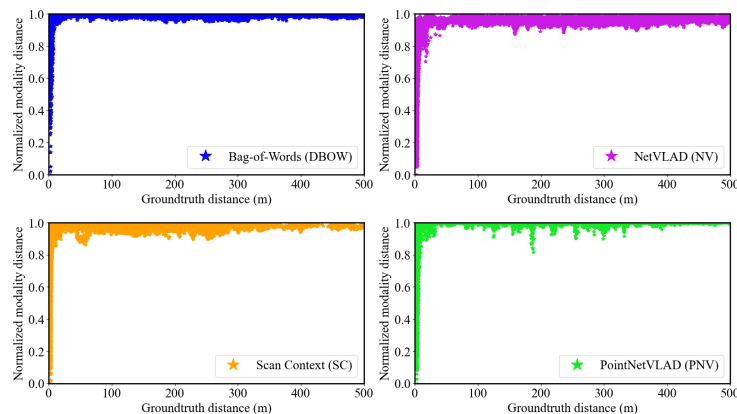




**Figure 4.** GNSS traces of our self-collected data in SATIE laboratory nearby cities (Saclay and Saint-Aubin).

### 5.2. Modalities Characterization

The performance of a place recognition method will highly depend on its ability to compute the similarity between the current frame descriptor and the loop closure descriptor. In principle, the model must incorporate probable measurement errors such as noise or blur and dynamic objects. These errors, being reflected in the computed descriptor, will result in incorrect loop detections (False Positive (FP)) or loop closure misses (False Negative (FN)). To avoid this, similarity measurement errors are modeled as Gaussians where the mean is the computed similarity and the standard deviation is an intrinsic noise parameter of the modality. In Figure 5, we illustrate the correlation between the ground truth distance (in meters) and Modality Distance between a current frame and a Loop Closure Candidate (LCC).



**Figure 5.** Normalized modality distance of each modality. Using all the frames of sequence 00 of the KITTI dataset, each plotted point represents the normalized modality distance in function of the ground truth distance of all possible pairs of frames in the sequence.

### 5.3. Optimal Particles Ratio

To validate the proposed approach using the Monte Carlo method and ensure its robustness, we selected sequence 00 from the KITTI dataset. For each of the 18 different particle-sampling ratio configurations, the performance was tested across the four modalities and evaluated over 50 executions. Each configuration is defined as follows:

- The first number represents the number of particles proposed based on the similarity derived from each modality.
- The second number represents the number of model-based particles generated around the proposed candidates for each perception modality.

For example, particle distribution 40\_200 means that 10 particles were proposed as potential LCC by each modality (using four modalities, that is 40 particles in total), and 50 particles were appended to the sampling pool based on the weight of the precedent candidates.



In what follows, we elaborate on the results obtained using these particle's ratios under different evaluation metrics commonly used in the SLAM context.

### 5.3.1. Place Recognition Metrics

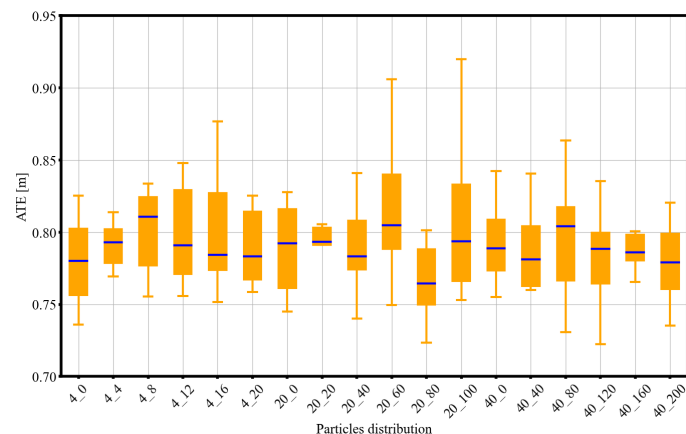
Table 2 provides a summary of the trial results, with the best outcomes highlighted in green. Based on the AUC, the maximum F1 score, and the Extended Precision (EP), the optimal configuration is 20\_40. Here, five particles are proposed by each modality and 10 particles are added based on a Gaussian model around the proposed particles.

**Table 2.** AUC, F1 score, and EP for various particle distributions on Seq. 00 of the KITTI dataset, with the best results highlighted in green.

Particles Distribution	AUC	F1	EP
4_0	0.95	0.94	0.94
4_4	0.96	0.94	0.95
4_8	0.96	0.94	0.95
4_12	0.95	0.94	0.94
4_16	0.95	0.94	0.94
4_20	0.96	0.94	0.95
20_0	0.94	0.95	0.96
20_20	0.94	0.95	0.96
20_40	0.96	0.95	0.96
20_60	0.96	0.94	0.96
20_80	0.96	0.95	0.95
20_100	0.94	0.95	0.95
40_0	0.94	0.95	0.96
40_40	0.94	0.94	0.95
40_80	0.94	0.95	0.96
40_120	0.94	0.95	0.96
40_160	0.94	0.95	0.95
40_200	0.93	0.95	0.95

### 5.3.2. ATE Evaluation

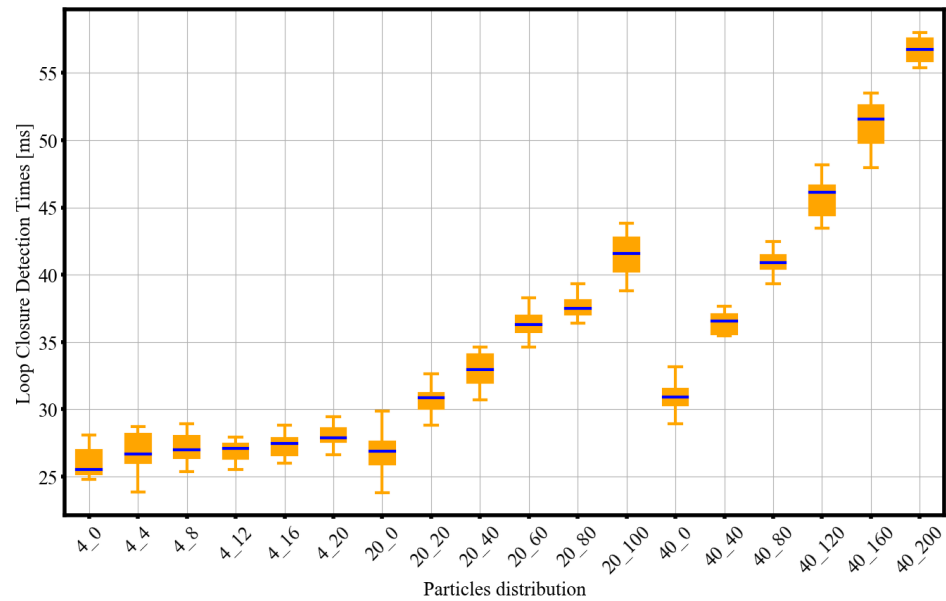
In Figure 6 we summarize the results of these trials. We find that the best configuration is the one given by 20\_80. In this configuration, five particles are proposed by each of the four modalities used and 20 particles are added to the pool of all considered candidates based on the Gaussian model around the similarity-proposed candidates. This configuration achieves a mean ATE of 0.76m (in contrast to the 20\_40 configuration, achieving a mean ATE of 0.78 m).



**Figure 6.** ATE (m) in Seq. 00 of KITTI dataset. On the x-axis, the first number represents the total number of LCCs suggested by the perception modalities, while the second number indicates the number of sampled loop candidates. The total is the sum of all considered loop candidates.

### 5.3.3. Timing Evaluation

In Figure 7 we summarize the results of these trials.



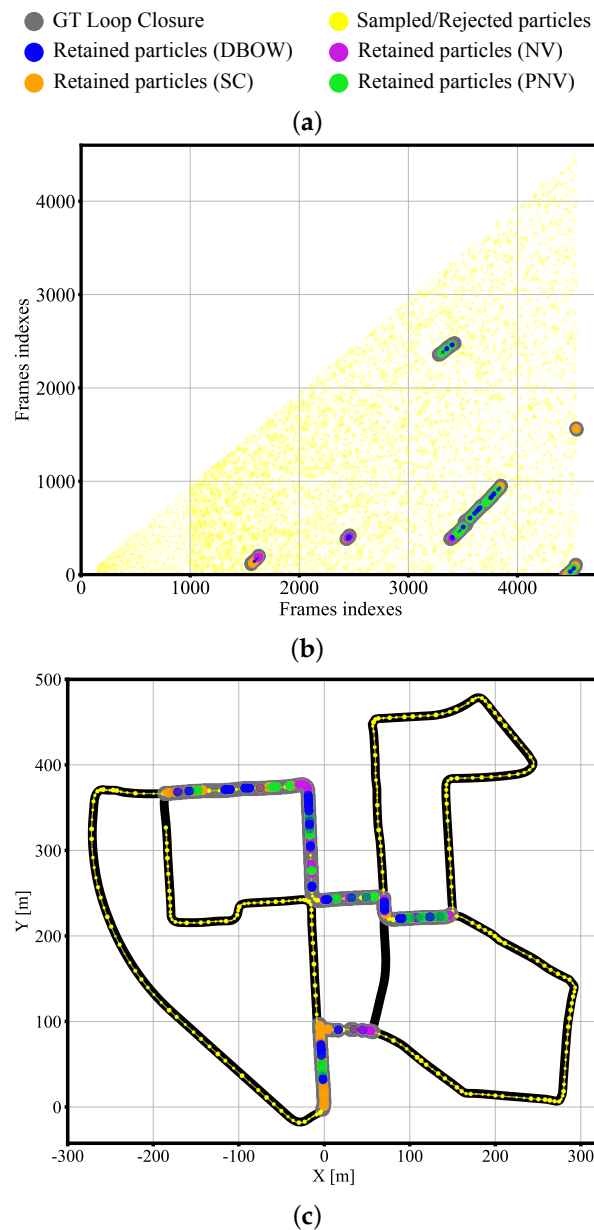
**Figure 7.** Processing time (ms) in Seq. 00 of KITTI dataset. On the x-axis, the first number represent the total number of LCCs proposed by the perception modalities. The second number indicates the number of sampled loop candidates. The sum being all the considered loop candidates.

Following these results, we found that a correct trade-off between performance in terms of Precision-Recall, trajectory accuracy and time consumption is the configuration 20\_80 where each modality can suggest five loop closure candidates, to which, 20 other candidates are sampled around these best five candidates. This enables the system to run at 256 Hz while guaranteeing good overall AUC, F1-maximum score and EP and an ATE at 0.76 m on average.

#### 5.4. Qualitative Evaluation

Building on the optimal particle ratio identified in the previous subsections through timing evaluations, localization errors, and place recognition metrics, we now turn our attention to validating the effectiveness of this configuration. This involves analyzing the distribution of sampled and rejected loop closure candidates along the trajectory and evaluating the contribution of each modality in detecting loop closures.

To validate the success of modalities modelization using the approach described in the previous subsection, we first show in Figure 8 how the sampled/rejected loop closure candidates are distributed over all previous frames of the current pose and how all modalities were able to contribute to detecting LC. This figure demonstrates how the proposed framework effectively generates probable LCCs across all traversed poses during the AGV's trajectory while eliminating inaccurate candidates (depicted as yellow dots). Conversely, the retained particles with the highest weights are concentrated within the ground truth loop closure area. We recall that a ground truth LC is defined as a pair of poses with a relative distance of less than 3 m.



**Figure 8.** Distributions of particles representing potential LCCs, identified by the four modalities over the Seq. 00 of KITTI dataset. (a) The gray particles indicate ground truth loop closures. The yellow particles are all the sampled particles that were dropped during the update step. The blue (magenta, orange, green) particles are based on DBOW (NetVLAD (NV), Scan Context (SC), PointNetVLAD (PNV)) similarity. (b) Distribution of particles across the sequence frames. (c) Distribution of particles across the sequence map. The black dots are the vehicle's poses.

### 5.5. Ablation Studies—Place Recognition Results

The results of the comparative experiments are demonstrated in Table 3. Compared to the state-of-the-art methods using multiple modalities for loop closure, MinkLoc++ and CORAL were evaluated only on the Seq. 00 of KITTI dataset and achieved an AR@1% (that is Average Recall taking into account 1% of the database size) of 82.1 and 76.4, respectively.

**Table 3.** Assessment of the AUC, F1 score, and EP for different particle distributions on Seq. 00 of the KITTI dataset. Cells with different colors classify total number of modalities (from one to four). Highlighted in green is the best result.

Methods	AUC						F1-Maximum Score						Extended Precision					
	00	02	05	06	07	08	00	02	05	06	07	08	00	02	05	06	07	08
1_D	0.88	0.70	0.79	0.91	0.67	0.29	0.85	0.79	0.73	0.93	0.78	0.42	0.84	0.72	0.70	0.80	0.78	0.42
1_N	0.92	0.77	0.79	0.98	0.83	0.19	0.87	0.85	0.86	0.98	0.88	0.28	0.77	0.74	0.88	0.96	0.88	0.28
1_S	0.92	0.83	0.82	0.99	0.83	0.78	0.88	0.88	0.80	0.99	0.87	0.75	0.87	0.77	0.80	0.99	0.87	0.75
1_P	0.77	0.53	0.29	0.97	0.56	0.36	0.81	0.58	0.33	0.97	0.77	0.45	0.81	0.56	0.50	0.97	0.77	0.45
2_D_N	0.93	0.80	0.81	0.98	0.82	0.34	0.94	0.86	0.77	0.98	0.87	0.46	0.91	0.79	0.76	0.98	0.87	0.46
2_S_P	0.94	0.81	0.83	0.99	0.83	0.82	0.95	0.87	0.81	0.99	0.88	0.79	0.95	0.79	0.81	0.99	0.88	0.79
2_D_P	0.92	0.72	0.74	0.98	0.73	0.61	0.93	0.80	0.74	0.98	0.80	0.66	0.80	0.75	0.75	0.98	0.80	0.66
2_N_P	0.92	0.71	0.79	0.99	0.83	0.42	0.91	0.82	0.86	0.99	0.90	0.50	0.92	0.72	0.88	0.99	0.90	0.50
2_D_S	0.94	0.95	0.82	0.99	0.83	0.92	0.94	0.96	0.82	0.99	0.88	0.84	0.95	0.75	0.84	0.99	0.88	0.84
2_N_S	0.94	0.87	0.85	0.99	0.83	0.79	0.95	0.88	0.81	0.99	0.87	0.77	0.92	0.88	0.80	0.99	0.87	0.77
3_D_N_P	0.93	0.82	0.74	0.99	0.82	0.51	0.94	0.86	0.72	0.99	0.87	0.59	0.87	0.77	0.74	0.99	0.87	0.59
3_D_N_S	0.92	0.96	0.83	0.99	0.83	0.67	0.93	0.96	0.83	0.99	0.88	0.70	0.86	0.82	0.85	0.99	0.88	0.70
3_D_S_P	0.94	0.89	0.81	0.99	0.82	0.67	0.94	0.91	0.82	0.99	0.87	0.70	0.92	0.81	0.84	0.99	0.87	0.70
3_N_S_P	0.93	0.87	0.82	0.99	0.86	0.78	0.94	0.88	0.80	0.99	0.90	0.75	0.95	0.88	0.83	0.99	0.90	0.75
4_D_N_S_P	0.94	0.96	0.86	0.99	0.87	0.94	0.97	0.96	0.84	0.99	0.92	0.87	0.97	0.79	0.86	0.99	0.92	0.87

Sensor configuration and environmental factors significantly affect the performance of multimodal learning-based techniques. Indeed, MinkLoc++, CORAL, NetVLAD and PointNetVLAD were all trained on the Oxford dataset, which differs considerably in characteristics from the KITTI dataset. Such dataset dependency, which is a known issue for learning-based techniques, results in a performance drop when applied to unseen scenarios and datasets with different characteristics (e.g., different lighting conditions, weather or sensor noise).

Even if the used modalities NetVLAD and PointNetVLAD were trained on the Robot Car Dataset, the obtained results using complementary modalities are better by a large margin. This is due to the fact that the proposed framework takes advantage also of geometric methods (DBOW and Scan Context) which demonstrate greater robustness when transitioning across datasets. Their performance tends to remain stable as they rely less on learned priors and more on intrinsic geometric relationships.

The presented multimodal results are grouped and classified by the number of modalities used and the information source. For example, 1\_D, 1\_N, 1\_S, 1\_P means that only one modality was used and that modality is DBOW, NetVLAD, Scan Context, or PointNetVLAD, respectively. The color classification refers to the combination strategy. In gray, modalities from the same source of information were combined (either camera (DBOW and NetVLAD) or LiDAR (Scan Context and PointNetVLAD)). In orange, two modalities from different sources of information are combined. In blue, three modalities are used and lastly, in red all four modalities are used.

Table 4 shows the same results of the proposed method validated on the self-collected dataset. Overall, for both tables, we find that combining multiple modalities can increase the performance of place recognition when considered as a classification problem. This is especially true when combining multiple modalities from different sensors.

**Table 4.** Analysis of the AUC, F1-maximum score, and Extended Precision across various ablation configurations on Self-Collected dataset. Cells with different colors classify total number of modalities (from one to four). Highlighted in green is the best result.

Methods	AUC			F1-Maximum Score			Extended Precision		
	01	02	03	01	02	03	01	02	03
1_D	0.79	0.99	0.79	0.85	0.99	0.83	0.70	0.99	0.81
1_N	0.72	0.99	0.72	0.81	0.99	0.82	0.81	0.99	0.85
1_S	0.69	0.80	0.51	0.80	0.87	0.60	0.83	0.71	0.54
1_P	0.58	0.39	0.69	0.69	0.62	0.81	0.67	0.53	0.66
2_D_N	0.80	0.99	0.93	0.87	0.99	0.93	0.89	0.99	0.93
2_S_P	0.69	0.89	0.72	0.82	0.93	0.76	0.73	0.72	0.63
2_D_P	0.81	0.99	0.81	0.88	0.99	0.84	0.61	0.99	0.81
2_N_P	0.80	0.91	0.75	0.83	0.93	0.82	0.83	0.93	0.84
2_D_S	0.79	0.99	0.83	0.86	0.99	0.85	0.82	0.99	0.70
2_N_S	0.74	0.99	0.74	0.83	0.99	0.81	0.80	0.99	0.84
3_D_N_P	0.80	0.99	0.89	0.87	0.99	0.89	0.85	0.99	0.85
3_D_N_S	0.86	0.99	0.86	0.90	0.99	0.89	0.87	0.99	0.89
3_D_S_P	0.82	0.99	0.81	0.86	0.99	0.83	0.87	0.99	0.69
3_N_S_P	0.80	0.96	0.72	0.89	0.97	0.75	0.85	0.88	0.77
4_D_N_S_P	0.87	0.99	0.83	0.92	0.99	0.87	0.93	0.99	0.88

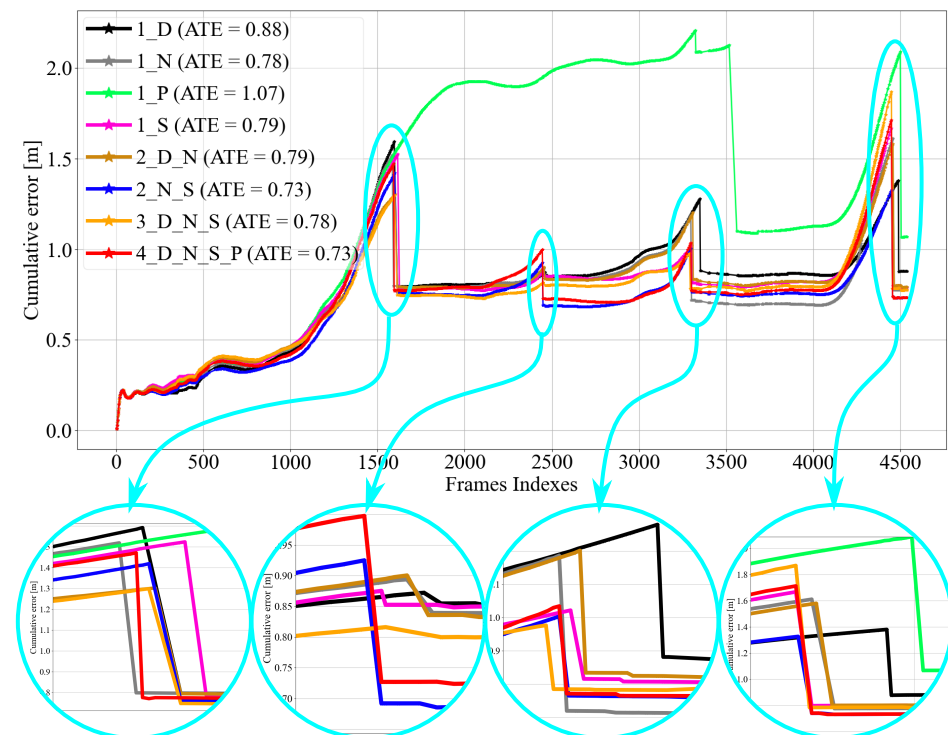
### 5.6. Ablation Studies—Cumulative Errors Results

Existing methods often evaluate SLAM systems at the end of a trajectory by calculating the ATE as introduced in [28], or the Relative Translation ( $t_{rel}$ ) and Rotation ( $r_{rel}$ ) errors described in [13]. However, an equally valuable assessment is the Cumulative Error observed during runtime. Figure 9 illustrates the progression of cumulative error during

the execution of ORB-SLAM2 [2], using its standard loop closure module based on DBOW (depicted in black). Other configurations for loop closure detection are represented in their respective colors.

We demonstrate that by implementing the proposed MMLC framework based on SGPF, loops are generally detected earlier than using only one modality, which leads to limiting the drift of the trajectory earlier. Finally, the cumulative error during runtime has been reduced by 51.90% (from 1.58m to 0.76m) in the first loop detected (comparing 1\_D which is the baseline ORB-SLAM2 without modification and 4\_D\_N\_S\_P). This result, emphasizes the importance of focusing on the loop closure detection functionality as well as the odometry block to further reduce the cumulative error and approach an accurate real-time SLAM system.

Indeed, incorporating four modalities (DBOW, SC, NV, PNV) significantly enhances the system's ability to detect the LC at the earliest occurrence, facilitating timely trajectory corrections and more effective mitigation of cumulative drift. This results in a 40.90% cumulative error reduction in the last loop detected when using the four modalities compared to using a single one. Table 5 summarizes the most relevant cumulative error reduction under various configurations of the MMLC.



**Figure 9.** Evolution of cumulative error during ORB-SLAM2 [2] runtime, using different modalities combinations for loop closure.



**Table 5.** Quantitative analysis of Figure 9 illustrating cumulative error improvement using different monomodal and multimodal loop closures executed on KITTI 00. Here, the baseline is 1\_D, corresponding to the classical ORB-SLAM2 framework, where only DBOW is used for loop closure. For each implementation, we highlight (in bold) the frame index of the earliest loop detected by a modality (third column). Then, we compute the cumulative error of each implementation at the same frame index of the earliest loop closure (fourth column). The last column is the error reduction (in green) or augmentation (in red) with respect to the baseline.

Revisited Area	LC Method/ Configuration	LC Frame Index	Error at the First LC Frame Index	Improvement (%)
1	1_D	1613	1.58	-
	1_N	1592	0.80	−49.36%
	1_S	1625	1.48	−6.33%
	1_P	-	1.52	−3.80%
	2_D_N	<b>1584</b>	<b>0.79</b>	−50%
	2_N_S	1586	1.40	−11.39%
	3_D_N_S	<b>1584</b>	<b>0.74</b>	−53.16%
	4_D_N_S_P	<b>1584</b>	<b>0.76</b>	−51.90%
2	1_D	2463	0.86	
	1_N	2461	0.89	+3.37%
	1_S	2450	0.87	+1.15%
	1_P	-	1.95	+55.90%
	2_D_N	2460	0.89	+3.37%
	2_N_S	<b>2449</b>	<b>0.69</b>	−19.77%
	3_D_N_S	2463	0.82	−4.65%
	4_D_N_S_P	<b>2449</b>	<b>0.72</b>	−16.28%
3	1_D	3352	1.18	-
	1_N	3301	1.18	0
	1_S	3310	1.01	−14.41%
	1_P	3326	2.17	+83.90%
	2_D_N	3311	1.17	−0.85%
	2_N_S	3302	1.00	−15.25%
	3_D_N_S	<b>3294</b>	<b>0.78</b>	−33.90%
	4_D_N_S_P	3301	1.03	−12.71%
4	1_D	4490	1.32	-
	1_N	4465	1.60	+21.21%
	1_S	4455	0.79	−40.15%
	1_P	4499	1.96	+48.48%
	2_D_N	4464	1.57	+18.94%
	2_N_S	4455	0.78	−40.90%
	3_D_N_S	<b>4454</b>	<b>0.78</b>	−40.90%
	4_D_N_S_P	4455	0.78	−40.90%

### 5.7. Ablation Studies—Final Trajectory Estimation Results

From the preceding analysis, we selected the 20\_80 configuration for distributing particles between similarity-based candidates and injected candidates (each modality can suggest five LCDs, to which, 20 other candidates are sampled around these best five candidates). Each sequence was executed 50 times, and the median value is reported in the presented results in Table 6 for the KITTI dataset and the Table 7 for the self-collected dataset.

The proposed framework consistently outperforms the traditional monomodal loop closure detection approach of ORB-SLAM2. This superiority is demonstrated using various metrics, including the Absolute Translation Error (ATE) introduced in [28] and the average relative translation ( $t_{rel}$ ) and rotation ( $r_{rel}$ ) errors presented in [13].

Sequence 08 of the KITTI dataset stands out as a notable example, as all loop closures occurred in the reverse direction. A camera-only place recognition method failed to detect these loop closures, leading to an ATE of 3.32 m. However, incorporating additional modalities that leverage LiDAR data successfully identified the loops despite the reversed revisitation. This reduced the accumulated error to 2.64 m, representing a 20.48% improvement.

Overall, by enhancing the SLAM system’s back-end with multimodal loop closure, without altering its front-end, the proposed framework achieved superior performance across all tested scenarios compared to the existing approach.

In order to rule out the possibility of one modality dominance, we tested the proposed method on our self-collected dataset where visual features are more salient than geometrical information used by Scan Context (SC) and on which neither NetVLAD (NV) nor PointNetVLAD (PNV) were trained. The findings were satisfactory since the proposed method resisted the False Positives proposed by SC and PNV. Finally, we reconfirm our observation, that is, combining multiple sources of information is more beneficial than combining multiple modalities using only one source of information. Indeed, the use of different sensors can offer a degree of redundancy, enhancing the system’s reliability. In situations where one modality may fail or be less effective, such as in low light conditions for visual sensors, other modalities can compensate, ensuring consistent performance across varying conditions.

**Table 6.** Assessment of the ATE, relative translation ( $t_{rel}$ ), and relative rotation ( $r_{rel}$ ) errors on the KITTI dataset, with the best result highlighted in green. Cells with different colors classify total number of modalities (from one to four).

Methods	ATE (m)						$t_{rel}$ (%)						$r_{rel}$ (°/100m)					
	00	02	05	06	07	08	00	02	05	06	07	08	00	02	05	06	07	08
1_D	0.85	3.94	0.76	0.83	0.38	3.32	0.72	0.77	0.44	0.53	0.52	1.08	0.25	0.24	0.19	0.16	0.29	0.32
1_N	0.78	3.94	0.45	0.76	<b>0.36</b>	3.52	0.71	0.77	0.43	<b>0.49</b>	<b>0.50</b>	1.07	0.25	0.24	0.17	0.15	<b>0.28</b>	0.32
1_S	0.78	3.56	0.35	0.72	0.37	2.65	0.71	0.78	<b>0.40</b>	<b>0.49</b>	<b>0.50</b>	1.06	0.25	0.24	<b>0.16</b>	<b>0.14</b>	<b>0.28</b>	0.31
1_P	0.98	7.77	1.61	0.78	0.38	3.21	0.73	0.83	0.59	0.52	0.51	1.05	0.25	0.27	0.24	0.15	<b>0.28</b>	0.32
2_D_N	0.77	3.78	0.39	0.75	0.37	3.28	0.71	0.77	0.41	0.52	0.51	1.06	0.25	0.24	0.17	0.15	0.29	0.31
2_S_P	0.77	3.69	0.34	0.74	0.37	2.65	0.71	0.79	<b>0.40</b>	0.53	<b>0.50</b>	<b>1.03</b>	0.25	0.24	<b>0.16</b>	0.16	<b>0.28</b>	<b>0.30</b>
2_D_P	0.81	3.94	0.74	0.77	0.38	3.12	0.71	0.78	0.45	0.50	0.52	<b>1.03</b>	0.25	0.24	0.19	0.15	0.29	<b>0.30</b>
2_N_P	0.78	3.94	0.48	0.74	0.37	3.20	<b>0.70</b>	0.77	0.43	0.51	<b>0.50</b>	1.05	0.25	0.24	0.17	<b>0.14</b>	<b>0.28</b>	0.32
2_D_S	0.77	<b>3.51</b>	0.35	0.71	0.37	<b>2.64</b>	<b>0.70</b>	0.78	<b>0.40</b>	0.51	0.51	1.04	0.25	<b>0.23</b>	<b>0.16</b>	<b>0.14</b>	<b>0.28</b>	0.31
2_N_S	<b>0.73</b>	3.54	0.33	<b>0.70</b>	0.37	2.65	<b>0.70</b>	<b>0.76</b>	<b>0.40</b>	0.51	<b>0.50</b>	<b>1.03</b>	0.25	<b>0.23</b>	<b>0.16</b>	0.15	<b>0.28</b>	0.31
3_D_N_P	0.77	3.76	0.49	0.77	0.37	3.14	<b>0.70</b>	0.77	0.42	0.51	0.52	1.04	0.25	0.24	0.17	0.15	0.29	<b>0.30</b>
3_D_N_S	0.75	<b>3.51</b>	0.34	0.74	0.37	2.70	<b>0.70</b>	0.77	<b>0.40</b>	0.51	0.51	1.04	0.25	0.24	<b>0.16</b>	0.15	0.29	<b>0.30</b>
3_D_S_P	0.77	3.53	0.35	0.76	0.37	2.67	0.71	0.77	<b>0.40</b>	0.52	0.51	<b>1.03</b>	0.25	0.24	<b>0.16</b>	<b>0.14</b>	0.29	<b>0.30</b>
3_N_S_P	0.75	3.54	0.36	0.73	0.37	2.65	<b>0.70</b>	0.77	<b>0.40</b>	0.53	<b>0.50</b>	<b>1.03</b>	0.25	0.24	<b>0.16</b>	0.15	<b>0.28</b>	0.31
4_D_N_S_P	<b>0.73</b>	<b>3.51</b>	<b>0.32</b>	0.71	<b>0.36</b>	<b>2.64</b>	<b>0.70</b>	0.77	<b>0.40</b>	0.50	0.51	1.05	0.25	0.24	<b>0.16</b>	<b>0.14</b>	<b>0.28</b>	0.32

To sum up, the use of four modalities (DBOW, SC, NV, PNV) statistically enhances the robustness and reliability of the system by leveraging their complementary strengths, resulting in significant improvements in trajectory estimation. We demonstrate that our proposed multimodal approach ensures greater availability of LCD possibility across diverse conditions, enabling consistent performance even in challenging environments. For

instance, in Sequence 00 of the KITTI dataset, the error is reduced by 14.12% and is reduced by 20.48% in sequence 08.

**Table 7.** Analysis of the ATE, relative translation ( $t_{rel}$ ), and relative rotation ( $r_{rel}$ ) errors on the self-collected dataset. Highlighted in green, the best result. Cells with different colors classify total number of modalities (from one to four).

Methods	ATE (m)			$t_{rel}$ (%)		
	01	02	03	01	02	03
1_D	4.38	2.25	1.66	1.67	1.01	0.93
1_N	4.49	2.37	1.73	1.66	1.05	0.98
1_S	4.58	2.45	1.79	1.72	1.09	1.00
1_P	4.80	2.73	1.77	1.76	1.15	0.99
2_D_N	4.38	2.23	1.60	1.64	1.03	0.91
2_S_P	4.53	2.46	1.76	1.73	1.10	0.97
2_D_P	4.33	2.22	1.65	1.64	0.91	0.92
2_N_P	4.35	2.37	1.73	1.61	1.05	0.99
2_D_S	4.39	2.22	1.64	1.64	1.01	0.92
2_N_S	4.49	2.28	1.73	1.69	1.03	0.97
3_D_N_P	4.38	2.18	1.63	1.66	1.02	0.92
3_D_N_S	4.29	2.23	1.65	1.61	1.02	0.93
3_D_S_P	4.31	2.29	1.68	1.63	1.02	0.94
3_N_S_P	4.35	2.28	1.73	1.64	1.03	0.97
4_D_N_S_P	4.29	2.21	1.64	1.60	0.99	0.91

## 6. Conclusions

In this extended study, we introduced a Multi-Modal Loop Closure (MMLC) detection framework designed for seamless integration into the loop closure module of any pose-graph optimization-based SLAM system. The proposed generalized methodology leverages N-modalities to enhance the accuracy of loop closure detection. To demonstrate the gain of combining information from multiple sources to infer the most probable loop closure, we used ORB-SLAM2 as the front end, incorporating both stereo and LiDAR frames within the loop closure module. The conducted experimental study focused on a four-modality extension, incorporating DBOW, NetVLAD, Scan Context and PointNetVLAD for processing camera and LiDAR data. This extension significantly enhanced the global optimization of pose estimates. The experimental results demonstrated that the proposed approach effectively minimizes the accumulated residual error at the earliest opportunity and can be integrated into a real-time SLAM system. Furthermore, the approach expands the functionality of loop closure (LC) operations while improving robustness against data unavailability and track loss scenarios. Additionally, it enhances the integrity of the SLAM system by identifying the most likely multimodal LC.

In our future research, we aim to develop a real-time multimodal loop closure (LC) module suitable for integration into an AGV, a project already initiated in [35,36]. The system targets frugal hardware architectures tailored for embedded systems to enhance computational efficiency. Additionally, we plan to expand the multimodal approach to the front-end by incorporating multiple information sources for odometry estimation.

**Author Contributions:** Conceptualization, M.C., S.R.F. and A.E.O.; methodology, M.C., S.R.F. and A.E.O.; software, M.C.; validation, M.C., S.R.F. and A.E.O.; formal analysis, M.C., S.R.F. and A.E.O.; investigation, M.C., S.R.F. and A.E.O.; resources, M.C., S.R.F. and A.E.O.; data curation, M.C., S.R.F. and A.E.O.; writing—original draft preparation, M.C.; writing—review and editing, S.R.F. and A.E.O.; visualization, M.C.; supervision, S.R.F. and A.E.O.; project administration, S.R.F. and A.E.O.; funding

acquisition, S.R.F. and A.E.O. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Restrictions apply to the datasets. The datasets presented in this article are not readily available because we are in the process of setting up a local server to publish the data, moreover the datasets are also part of another ongoing study. Requests to access the datasets should be directed to Sergio Rodríguez Flórez (sergio.rodriguez@universite-paris-saclay.fr).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Arshad, S.; Kim, G.W. Role of deep learning in loop closure detection for visual and lidar SLAM: A survey. *Sensors* **2021**, *21*, 1243.
2. Mur-Artal, R.; Tardós, J.D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262.
3. Gálvez-López, D.; Tardos, J.D. Bags of binary words for fast place recognition in image sequences. *IEEE Trans. Robot.* **2012**, *28*, 1188–1197.
4. Kim, G.; Kim, A. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In Proceedings of the 2018 IEEE/RISJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 4802–4809.
5. Wang, W.; Min, H.; Wu, X.; Hou, X.; Li, Y.; Zhao, X. High Accuracy and Low Complexity LiDAR Place Recognition using Unitary Invariant Frobenius Norm. *IEEE Sens. J.* **2022**, *23*, 11205–11217.
6. Chghaf, M.; Rodriguez, S.; Ouardi, A.E. Camera, LiDAR and multi-modal SLAM systems for autonomous ground vehicles: A survey. *J. Intell. Robot. Syst.* **2022**, *105*, 1–35.
7. Cattaneo, D.; Vaghi, M.; Fontana, S.; Ballardini, A.L.; Sorrenti, D.G. Global visual localization in LiDAR-maps through shared 2D-3D embedding space. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation, Paris, France, 31 May–31 August 2020; pp. 4365–4371.
8. Komorowski, J.; Wysoczańska, M.; Trzcinski, T. MinkLoc++: Lidar and monocular image fusion for place recognition. In Proceedings of the 2021 International Joint Conference on Neural Networks, Shenzhen, China, 18–22 July 2021; pp. 1–8.
9. Pan, Y.; Xu, X.; Li, W.; Cui, Y.; Wang, Y.; Xiong, R. Coral: Colored structural representation for bi-modal place recognition. In Proceedings of the 2021 IEEE/RISJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 2084–2091.
10. Yu, X.; Zhou, B.; Chang, Z.; Qian, K.; Fang, F. MMDF: Multi-Modal Deep Feature Based Place Recognition of Mobile Robots with Applications on Cross-scene Navigation. *IEEE Robot. Autom. Lett.* **2022**, *7*, 6742–6749.
11. Chghaf, M.; Flórez, S.R.; El Ouardi, A. A multimodal loop closure fusion for autonomous vehicles SLAM. *Robot. Auton. Syst.* **2023**, *165*, 104446.
12. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. The revisiting problem in simultaneous localization and mapping. In *Online Appearance-Based Place Recognition and Mapping: Their Role in Autonomous Navigation*; Springer: Cham, Switzerland, 2022; pp. 1–33.
13. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
14. Williams, B.; Cummins, M.; Neira, J.; Newman, P.; Reid, I.; Tardós, J. A comparison of loop closing techniques in monocular SLAM. *Robot. Auton. Syst.* **2009**, *57*, 1188–1197.
15. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5297–5307.
16. Zhang, Y.; Liu, R.; Yu, H.; Zhou, B.; Qian, K. Visual Loop Closure Detection With Instance Segmentation and Image Inpainting in Dynamic Scenes Using Wearable Camera. *IEEE Sens. J.* **2022**, *22*, 16628–16637.
17. Xiao, D.; Li, S.; Xuanyuan, Z. Semantic Loop Closure Detection for Intelligent Vehicles Using Panoramas. *IEEE Trans. Intell. Veh.* **2023**, *8*, 4395–4405.
18. Uy, M.A.; Lee, G.H. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–2 June 2018; pp. 4470–4479.
19. Guo, H.; Zhu, J.; Chen, Y. E-LOAM: LiDAR odometry and mapping with expanded local structural information. *IEEE Trans. Intell. Veh.* **2022**, *8*, 1911–1921.
20. Xiang, H.; Zhu, X.; Shi, W.; Fan, W.; Chen, P.; Bao, S. DeLightLCD: A Deep and Lightweight Network for Loop Closure Detection in LiDAR SLAM. *IEEE Sens. J.* **2022**, *22*, 20761–20772.

21. Piasco, N.; Sidibé, D.; Demonceaux, C.; Gouet-Brunet, V. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognit.* **2018**, *74*, 90–109.
22. Collier, J.; Se, S.; Kotamraju, V. Multi-sensor appearance-based place recognition. In Proceedings of the 2013 International Conference on Computer and Robot Vision, Regina, SK, Canada, 28–31 May 2013; pp. 128–135.
23. Yuan, Z.; Xu, K.; Zhou, X.; Deng, B.; Ma, Y. SVG-Loop: Semantic–Visual–Geometric Information-Based Loop Closure Detection. *Remote Sens.* **2021**, *13*, 3520.
24. Peng, T.; Zhang, D.; Liu, R.; Asari, V.K.; Loomis, J.S. Evaluating the power efficiency of visual SLAM on embedded GPU systems. In Proceedings of the 2019 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, USA, 15–19 July 2019; pp. 117–121.
25. Nguyen, D.D.; Elouardi, A.; Florez, S.A.R.; Bouaziz, S. HOOFR SLAM system: An embedded vision SLAM algorithm and its hardware-software mapping-based intelligent vehicles applications. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 4103–4118.
26. Manning, C.D. *Introduction to Information Retrieval*; Syngress Publishing: Rockland, MA, USA, 2008.
27. Ferrarini, B.; Waheed, M.; Waheed, S.; Ehsan, S.; Milford, M.J.; McDonald-Maier, K.D. Exploring performance bounds of visual place recognition using extended precision. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1688–1695.
28. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 573–580.
29. Kümmerle, R.; Steder, B.; Dornhege, C.; Ruhnke, M.; Grisetti, G.; Stachniss, C.; Kleiner, A. On measuring the accuracy of SLAM algorithms. *Auton. Robot.* **2009**, *27*, 387–407.
30. Ferrarini, B. Improving Visual Place Recognition in Changing Environments. Ph.D. Thesis, University of Essex, Essex, UK, 2023.
31. Chang, W.Y.; Chen, C.S.; Jian, Y.D. Visual tracking in high-dimensional state space by appearance-guided particle filtering. *IEEE Trans. Image Process.* **2008**, *17*, 1154–1167.
32. Pérez, J.; Caballero, F.; Merino, L. Enhanced monte carlo localization with visual place recognition for robust robot localization. *J. Intell. Robot. Syst.* **2015**, *80*, 641–656.
33. Djuric, P.M.; Kotecha, J.H.; Zhang, J.; Huang, Y.; Ghirmai, T.; Bugallo, M.F.; Miguez, J. Particle filtering. *IEEE Signal Process. Mag.* **2003**, *20*, 19–38.
34. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. *Int. J. Robot. Res. (IJRR)* **2013**, *32*, 1231–1237.
35. Chghaf, M.; Florez, S.A.R.; El Ouardi, A. A stereo vision geometric descriptor for place recognition and its GPU acceleration for autonomous vehicles applications. In Proceedings of the XXVIIIème Colloque Francophone de Traitement du Signal et des Images, GRETSI'22, Nancy, France, 6–9 September 2022.
36. Chghaf, M.; Flórez, S.R.; El Ouardi, A.; Bouaziz, S. Real-time embedded large-scale place recognition for autonomous ground vehicles using a spatial descriptor. In Proceedings of the Real-Time Processing of Image, Depth and Video Information 2023, Prague, Czech Republic, 24–28 April 2023; Volume 12571, pp. 11–21.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.