ICINCO 2025

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

PROCEEDINGS

Volume 2

Marbella, Spain

20 - 22 October, 2025

EDITORS

Giuseppina Carla Gini
Radu-Emil Precup
Dimitar Filev

https://icinco.scitevents.org

SPONSORED BY

INSTICC

PAPERS AVAILABLE AT

SCITEPRESS DIGITAL LIBRARY

ICINCO 2025

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

Volume 2

Marbella - Spain

October 20 - 22, 2025

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

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

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

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

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

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

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

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

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

Towards Transparent, Physically Consistent Machine Learning Models
Robert Babuska 5

Dealing with “Dirty” Data: Solutions from Fuzzy Systems Research
Uzay Kaymak 7

Data Aware Agentic AI: Cognitive, Control, and Data Flows
Michael Berthold 9

ROBOTICS AND AUTOMATION

FULL PAPERS

Sensorless Admittance Control of a Manipulator Arm Using a Nonlinear Observer for Force and Velocity Estimation
Brahim Brahmi 15

ALEX-GYM-1: A Novel Dataset and Hybrid 3D Pose Vision Model for Automated Exercise Evaluation
Ahmed Hassan, Abdelaziz Serour, Ahmed Gamea and Walid Gomaa 24

Evaluation of YOLO Architectures for Automated Transmission Tower Inspection Under Edge Computing Constraints
Gabriel Jose Scheid, Ronnier Frates Rohrich and André Schneider de Oliveira 35

A Depth Image Processing Algorithm for Monitoring Product Flow on a Conical Feeder Unit of a Multihead Weigher
Julia Isabel Hartmann and Christoph Ament 43

Towards Guaranteed Collision Avoidance for Multiple Autonomous Underactuated Unmanned Surface Vehicles in Restricted Waters
Erick J. Rodríguez-Seda 53

High-Precision Contour Tracking for Mobile Manipulators in Large-Scale Industrial Applications
Buu Hai Dang Trinh, Daniel Heß and Christof Röhrig 60

Δ-Y Transformations in Manipulator’s Stiffness Analysis
Alexandr Klimchik and Anatol Pashkevich 72

CSDF-by-SIREN: Learning Signed Distances in the Configuration Space Through Sinusoidal Representation Networks
Christoforos Vlachos and Konstantinos Moustakas 82

Pressing Force Regulation in Robotic 3D Printing via CFD-Aided Nozzle Posture Control
Shinichi Ishikawa, Ryo Yamada, Wakana Tsuru and Ryosuke Tasaki 91

A Digital Twin Enabled Runtime Analysis and Mitigation for Autonomous Robots Under Uncertainties
Jalil Boudjadar and Mirgita Frasheri 96

HOI-LCD: Leveraging Humans as Dynamic Landmarks Toward Thermal Loop Closing Even in Complete Darkness 105
Tatsuro Sakai, Yanshuo Bai, Kanji Tanaka, Wuhao Xie, Jonathan Tay Yu Liang and Daiki Iwata

Design Space Exploration and Performance Evaluation of a Multi-Chamber, Multi-Curvature Soft Actuator for Robotic Applications 116
Ansari Usama and Asokan Thondiyath

Towards Fully Onboard State Estimation and Trajectory Tracking for UAVs with Suspended Payloads 128
Martin Jiroušek, Tomáš Báča and Martin Saska

NAMOUnc: Navigation Among Movable Obstacles with Decision Making on Uncertainty Interval 139
Kai Zhang, Eric Lucet, Julien Alexandre Dit Sandretto, Shoubin Chen and David Filliat

A Web-Controlled, Modular 3D-Printed Exoskeleton for Upper Limb Stroke Recovery 150
Crina Bărbieru and Isabela Drămnesc

Redundancy Resolution in Multiple Feasibility Maps via MultiFM-RRT 158
Marc Fabregat-Jaén, Adrián Peidró, María Flores, Luis Payá and Óscar Reinoso

Place Recognition Using Bag of Semantic and Visual Words from Equirectangular Images 166
María Flores, Marc Fabregat-Jaén, Juan José Cabrera, Adrián Peidró, David Valiente and Luis Payá

Method for Automated Forklift Pallet Transfer with Simple Camera Calibration 175
Tibor Bataljak Savić, Krešimir Turković, Ian Petek and Damjan Miklić

SHORT PAPERS

ORB-Based Map Fusion with Position Transformation for Enhanced Pairwise Connection 185
Lucas Alexandre Zick, Dieisson Martinelli, Andre Schneider de Oliveira and Vivian Cremer Kalempa

A Scalable Robot-Agnostic Voice Control Framework for Multi-Robot Systems 194
Valentina Pericu, Federico Rollo and Navvab Kashiri

CLIP-LLM: A Framework for Autonomous Plant Disease Management in Greenhouse 202
Muhammad Salman, Muhayy Ud Din and Irfan Hussain

Point Cloud Registration for Visual Geo-Referenced Localization Between Aerial and Ground Robots 211
Gonzalo Garcia and Azim Eskandarian

Design and Control of a New Wrist Rehabilitation Robot 219
Simona-Daiana Stiole, Pusca Alexandru, Paul Tucan, Iuliu Nadas, Vasile Bulbucan, Andrei Cailean, Dragos Sebeni, Alexandru Banica, Daniela Jucan, Radu Morariu, Calin Vaida, Petru Dobra, Jose Machado and Doina Pisla

FOPID-Based Trajectory Control for an Unmanned Aerial Robotic Manipulator 227
Gabriela M. Andaluz, Zahid Nazate, Paulo Leica and Guillermo Palacios-Navarro

Towards Universal Detection and Localization of Mating Parts in Robotics 235
Stefan Marx, Attique Bashir and Rainer Müller

Development of a Control System for an Innovative Parallel Robot Used in Laparoscopic Pancreatic Surgery 243
Doina Pisla, Andra Ciocan, Bogdan Gherman, Diana Schlanger, Alexandru Pusca, Nadim Al Hajjar, Emil Mois, Andrei Cailean, Nicoleta Pop, Calin Vaida, Paul Tucan and Ionut Zima

Model Predictive Control for Crowd Navigation via Learning-Based Trajectory Prediction
Mohamed Parvez Aslam, Bojan Derajic, Mohamed-Khalil Bouzidi, Sebastian Bernhard and Jan Oliver Ringert 251

Comparative Analysis of Robotic Topologies for Transmission Line Inspection
Davi Riiti Goto Valle, Ronnier Frates Rohrich and André Schneider de Oliveira 259

ROSBLOCKS: A Visual Programming Interface for ROS2 Robots
Fernando Costa Nogueira, Dieisson Martinelli, Lucas Alexandre Zick, André Schneider de Oliveira and Vivian Cremer Kalempa 267

Predicting Contact Surfaces in Repetitive Robotic Tasks
Luis Hernán Campos, José Luis Reyes Ramos, Marcelo Fajardo-Pruna, Christian Tutivén and Carlos Saldarriaga 275

Simultaneous Learning of State-to-State Minimum-Time Planning and Control
Swati Dantu, Robert Pěnička and Martin Saska 283

Real-Time Hand Gesture Control of a Robotic Arm with Programmable Motion Memory
Daniel Giraldi Michels, Davi Giraldi Michels, Lucas Alexandre Zick, Dieisson Martinelli, André Schneider de Oliveira and Vivian Cremer Kalempa 292

A New Sliding Mode Control Proposal with a Clegg Integrator for a Mobile Manipulator
Pablo Proaño, Paulo Leica and Gabriela Andaluz 301

Optimizing Collision Avoidance in Dynamic Multi-Robot Systems: A Velocity Obstacle and BB-PSO Approach with Priority Consideration
Luis H. Sanchez-Vaca, Gildardo Sanchez-Ante and Hernan Abaunza 309

Curvature-Constrained Motion Planning and Control for Traffic Cone Manipulation Robot
Rudolf Krecht and Áron Ballagi 317

Calibration Architecture for the Nonlinear Wheel Odometry Model with Integrated Noise Compensation
Máté Fazekas and Péter Gáspár 325

3D Hand-Eye Calibration for Collaborative Robot Arm: Look at Robot Base Once
Leihui Li, Lixuepiao Wan, Volker Krueger and Xuping Zhang 333

Intelligent Surveillance System Using Deep Learning to Reduce Shoplifting in Minimarkets in Santiago de Surco, Lima, Peru
Yosep Alexeis Solorzano Aguero and Jose Karim Candela Rengifo 342

Manipulability Maximization of a Liquid-Handling Manipulator for Sloshing Suppression via Container Tilting
Ryuji Nakagawa and Ryosuke Tasaki 350

Simulation-Driven Design and Optimization of a Parametric Flat-Foot with Elastic Pads for a Planar Biped Robot
Koray Kadir Şafak and Oğuzhan Aykut Ekşioğlu 357

Design and Implementation of a Robotized Laser Module for Weed Control
Kai Blanco, Luis Emmi and Roemi Fernández 363

Experimental Validation of Load Attitude Estimation Using Computer Vision and IMU-Based Approaches for Slung-Load Aerial Robots 371
Shlok Panchal, Barbie Sharma, Yash Dadheech, Darshil Shah, Ayush Agnihotri, Kalash Jain, Parth S. Thakar and Anilkumar Markana

Hematoxylin and Eosin Stained Histopathological Image Enhancement Method 378
Bogusław Cyganek

Robust Scene Understanding for Mobile Robots Based on Vision and Deep Learning Models 386
Leticia C. Pereira and Fernando S. Osório

Augmenting Neural Networks-Based Model Approximators in Robotic Force-Tracking Tasks 394
Kevin Saad, Vincenzo Petrone, Enrico Ferrentino, Pasquale Chiacchio, Francesco Braghin and Loris Roveda

Leveraging ROS to Support LLM-Based Human-Robot Interaction 402
Walleed Khan, Deeksha Chandola, Enas AlTarawenah, Baran Parsai, Ishan Mangrota and Michael Jenkin

Kinematic and Dynamic Analysis of Quadruped Legged Robots: A New Formulation Approach 410
Vyshak Sureshkumar, Khalifa H. Harib and Adewale Oriyomi Oseni

Fire-Resistant Wall-Climbing UAV for Victim Detection in Urban Search and Rescue Missions 417
Vedant Hambire, Harsh Yadav, Dipshikha Hazari and Satyam Singh

INDUSTRIAL INFORMATICS

FULL PAPERS

An Efficient and Scalable Hyperdimensional Computing Framework for Anomaly Classification in Industrial Systems 429
Víctor Ortega, Soledad Escolar, Fernando Rincón, Jesús Barba and Julián Caba

Semi-Supervised Anomaly Detection in Directed Energy Deposition Using Thermal Images 437
Ufuk Ismail Ozdek, Yigit Kaan Tonkaz, Shawqi Mohammed Farea and Mustafa Unel

Real-Time Automated Visual Inspection of Decorative Wood Panels for Zero Defects Manufacturing 446
Beatriz Coutinho, Tomás Martins, Eliseu Pereira and Gil Gonçalves

Improving Industrial Interoperability and Scalability Through OPC-UA and Smart Object-Based Architectures 457
Guilherme Coelho, Liliana Antão, Beatriz Coutinho, Gil Gonçalves, António Augusto and Miguel Moura

SHORT PAPERS

Safety-Centric Monitoring of Structural Configurations in Outdoor Warehouse Using an UAV
Assia Belbachir, Antonio M. Ortiz, Ahmed Nabil Belbachir and Emanuele Ciccia 471

Noise-Robust Speech Transcription with Quantized Language Model Correction for Industrial Settings
Marco Murgia, Marco Fontana, Alberto Pes, Diego Reforgiato Recupero, Giuseppe Scarpi and Leonardo Daniele Scintilla 479

Fault Diagnosis of Industrial Robots Using a Digital Twin and GRU-Based Deep Learning
Ilhem Ben Hnaien, Eric Gascard, Zineb Simeu-Abazi and Hedi Dhouibi 487

Towards Industry 5.0: AAS/MLOps-Driven Model Maintenance for Data-Centric Production
Kiavash Fathi, Marcin Sadurski, Stefan Waskow, Tobias Kleinert and Hans Wernher van de Venn 495

Anomaly Detection in Directed Energy Deposition: A Comparative Study of Supervised and Unsupervised Machine Learning Algorithms
Berke Ayyıldızlı, Beyza Balota, Kerem Tatari, Shawqi Mohammed Farea and Mustafa Unel 503

Smart Optimized Scheduling Under Constraints in Industry 5.0 Through Intelligent Computational Methods
Cherifa Nakkach, Wiem Abbes and Yvan Picaud 511

Leveraging Edge and Fog Resources While Complying with EU's GDPR
Matilde Silva, Pedro C. Diniz and Gil Gonçalves 518

Digital Transformation of the Nuclear Industry: Leveraging Robotics, AI, and Digital Twins for Standardised, Safe, and Efficient Operations
Abdenour Benkrid, Omar Zahra, Réka Szõke, Ankur Shukla and István Szõke 526

SPECIAL SESSION ON BRIDGING THE GAP IN COLLABORATIVE ROBOTICS: FROM THEORY TO REAL APPLICATIONS

FULL PAPERS

Preliminary Design and Control of an Operator-Assistance System Integrated into a Cobot, for Anatomical Meat-Cutting Process
Alexis Babut, Chedli Bouzgarrou, Laurent Sabourin and Nicolas Bouton 541

Design and Validation of Sensorized Tools for Deformable Object Manipulation in Meat Cutting and Doll Demoulding
Saltanat Seitzhan, Dionisio Cartagena González, Alexis Babut, Daniel Sánchez-Martínez, Juan Antonio Micó, Chedli Bouzgarrou and Juan Antonio Corrales Ramón 552

SHORT PAPERS

Soft Robotics for Advanced Handling of Delicate Fruit Products
Luan Lang, Rodrigo Antunes, Martim Lima de Aguiar, Nuno Pereira, Thiago Assis Dutra, Yebo Lu and Pedro Dinis Gaspar 569

Automated Computational Workflow for the Parametric Design and Optimization of a 3D-Printed Fin Ray Effect Soft Robotic Finger
Rodrigo Antunes, Luan Lang, Martim Lima de Aguiar, Nuno José Matos Pereira, Thiago Assis Dutra, Yebo Lu and Pedro Dinis Gaspar 577

Differential Kinematics Control Using Circles as Bivectors of Conformal Geometric Algebra
Julio Zamora-Esquivel, Alberto Jaimes Pita, Edgar Macias-Garcia, Javier Felip-Leon, David Gonzalez-Aguirre and Eduardo Bayro-Corrochano 585

SPECIAL SESSION ON TRUSTWORTHY AND INTELLIGENT SMART AGRICULTURE SYSTEMS: AI, BLOCKCHAIN, AND IOT CONVERGENCE

FULL PAPERS

Integrated Approaches to Monitoring GIAHS Territories: Requirements, Telematics, Sensorization and Intelligent Management Solutions
Joel Soares, Carlos Teixeira and Ramiro Gonçalves 597

Smart Urban Tree Valorization: An AI-Blockchain-Based Application for the Preservation of Remarkable Trees
Hajer Nabli, Issra Jegham, Yasmine Zorgati, Rania Ajmi, Raoudha Ben Djemaa and Layth Sliman 609

AgriTrustChain: A Decentralized Certification and Edaphic Data Traceability Framework with Zero-Leak for Sustainable Farming Using Blockchain
Wafa Ben Slama Souei, Mohamed Amine Hattab, Layth Sliman, Raoudha Ben Djemaa and Faiza Khebour Allouche 621

Witness Byzantine Fault Tolerance with Signature Tree and Proof-of-Navigation for Wide Area Visual Navigation
Nasim Paykari, Taylor Clark, Ademi Zain, Damian Lyons and Mohamed Rahouti 633

SHORT PAPER

Design and Field Evaluation of a Robotic Cotton Harvester with Improved Structural Balance and Suction Mechanism
Van Patiluna, Joe Mari Maja, Aashish Karki and Edward Barnes 647

AUTHOR INDEX 655

Place Recognition Using Bag of Semantic and Visual Words from Equirectangular Images

María Flores[a], Marc Fabregat-Jaén[b], Juan José Cabrera[c], Adrián Peidró[d], David Valiente[e] and Luis Payá[f]

Engineering Research Institute of Elche (I3E), Miguel Hernández University, Avda. de la Universidad, s/, 03202, Elche, Alicante, Spain

{*m.flores, mfabregat, juan.cabreram, apeidro, dvaliente, lpaya*}*@umh.es*

Keywords: Place Recognition, Equirectangular Images, Semantic Information, Bag of Visual Words.

Abstract: Place recognition has a crucial relevance in some tasks of mobile robot navigation. For example, it is used for the detection of loop-closure or for estimating the position of a mobile robot along a route in a known environment. If place recognition is based on visual information, it can be approached as an image retrieval problem. The Bag of Visual Words technique can be used for image retrieval. Image retrieval is based on an image representation (for example, a vector) that contains relevant visual information. In this paper, two image signatures are proposed. Both are based on semantic and visual information. A bag of visual words is created for each semantic class. Local feature descriptors are classified according to the projection of their associated point on a segmented semantic map. On the one hand, the image signature is composed of a set of histograms where each cell encodes the frequency with which a visual word appears in the image. On the other hand, the image signature is composed of a set of vectors where each cell encodes the sum of the cosine distance between the visual word and the nearest extracted features.

1 INTRODUCTION

A mobile robot can navigate autonomously in a priori unknown environment or, by contrast, in a known environment. In the first situation, the mobile robot must solve the Simultaneous Localization And Mapping (SLAM) problem. This means that the mobile robot builds a map of the environment and simultaneously estimates its position within the map during navigation. In visual SLAM, the accuracy of the map and localization can be improved by identifying a previously visited location (Loop Closure Detection module). In the second situation, the map is already available in advance, and the mobile robot must estimate its actual position within this map. In this context, the mobile robot can locate itself if it is able to identify its current surroundings on the stored topological map.

Place recognition is a computer vision task in which, given an image, its location is identified by querying the locations of images which belong to the same place in a large geotagged database (Zeng et al., 2018). It is commonly posed as an image retrieval task. Image retrieval techniques can be grouped into the following: Text-Based Image Retrieval (TBIR) and Content-Based Image Retrieval (CBIR). The main difference between both is that the search for images similar to a given query from a large database is based on their visual content (CBIR) or on the textual data (metadata) associated with the image. The CBIR is based on three key components: selection, extraction, and representation of features. A comprehensive survey of this is presented by Srivastava et al. (2023). Similarly, Li et al. (2021) provide another survey of the fast advances and applications of theories and algorithms, focusing on those within the period from 2009 to 2019.

In place recognition, an important issue is the sensors on board the mobile robot, since they provide information about its surrounding, and its location is identified by analysing this captured information. Some well-recognized types of sensors used in place recognition are vision (Wang et al., 2018; Xu et al., 2019; Alfaro et al., 2024) and LiDAR systems (Cabr-

[a] https://orcid.org/0000-0003-1117-0868
[b] https://orcid.org/0009-0002-4327-0900
[c] https://orcid.org/0000-0002-7141-7802
[d] https://orcid.org/0000-0002-4565-496X
[e] https://orcid.org/0000-0002-2245-0542
[f] https://orcid.org/0000-0002-3045-4316

Flores, M., Fabregat-Jaén, M., Cabrera, J. J., Peidró, A., Valiente, D. and Payá, L.
Place Recognition Using Bag of Semantic and Visual Words from Equirectangular Images.

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 2*, pages 166-174
ISBN: 978-989-758-770-2; ISSN: 2184-2809

era et al., 2024; Vilella-Cantos et al., 2025).

This work focuses on solving place recognition using visual information captured in images. An important feature of images is the rich information they provide about the environment in which they were taken. In mobile robotics, the field of view of the vision system is relevant, since, for example, the wider the field of view, the fewer images are needed to create a map.

The present work is based on (Ouni et al., 2022). The authors propose three image signatures in order to resolve the image retrieval task. Focusing specifically on the type of signature that combines visual features and semantic information in (Ouni et al., 2022), this paper evaluates its behaviour for place recognition and in images with a wider field of view such as equirectangular images. This signature is an NxM matrix where N is the number of classes and M is the size of the visual descriptor. The procedure consists in extracting local feature points and a semantic segmented map, the points will be classified according to their projection on such map so that a semantic label is assigned. In this way, there are different clusters (one per semantic class) composed of visual descriptors. Each row of the image signature will be the centroid of each of these clusters. However, it may happen that a class has a completely different visual appearance. For this reason, we propose to create a bag of visual words for each semantic class instead of a class being represented by a unique visual descriptor.

The contributions of this work are as follows:

1. An image signature in which semantic and visual information is merged. Each semantic category will have a bag of visual words that will be used to obtain a frequency histogram. The signature will be the concatenation of all these histograms.
2. The frequency histograms of the previous signature are replaced by one-dimensional vectors, where each bin represents the sum of the cosine distances to each visual word. In other words, image encoding is based on distance instead of frequency.
3. The above image signatures and one of the proposal in (Ouni et al., 2022) (BoSW) are compared for place recognition in an outdoor environment. We want to clarify that we have only evaluated the image signature that the authors describe in Section 3.1 (denominated BoSW) of their paper, not their full global framework for CBIR.
4. These image signatures are evaluated in image retrieval when images are distorted, such as in equirectangular images.
5. The influence of the type of distance on the Nearest Neighbour search in the image encoding step is studied.

The remainder of this paper is organized as follows. In Section 2, some image retrieval techniques proposed in the literature are presented. Section 3 describes the different parts of the algorithm employed to solve image retrieval in this paper. Section 4 is focused on the experimental part, the database used, and the analytical metrics are described. The results obtained from the experiments are presented and discussed in Section 5. Finally, Section 6 presents the conclusions.

2 RELATED WORKS

2.1 Image Retrieval

In image retrieval, there are approaches based on bag of visual words. Mansoori et al. (2013) focus on the feature extraction stage and propose to incorporate colour information (hue descriptor) in descriptor features (SIFT) of the images.

In terms of works that propose combining semantic information and local visual features, Ouni et al. (2022) present three procedures in order to construct the image signatures using semantic information. Amongst the three signature, only two combine these two types of information, the other is based on semantic information only. One year later, Ouni et al. (2023) proposed two additional types of signatures. The first of these integrates at the same time the semantic proportions of objects and their spatial positions. Meanwhile the second one builds a semantic bag of visual phrase (i.e. a set of words linked together) by combining the visual vocabulary with semantic information. In this case, the image signature is an upper triangular matrix whose height and width are equal to the number of visual words in the codebook.

As with other computer vision tasks, the use of convolutional neural network-based architectures has increased in popularity over the last decade. Rani et al. (2025) propose a separable convolutional neural networks-based framework. This contribution reduces the computational complexity in terms of a number of convolutional operations and hyperparameters. Forcen et al. (2020) present a new representation of co-occurrences from deep convolutional features which is combined with the feature map in order to improve the image representation. Dubey (2022) presents a comprehensive survey of deep learning-based progress over the last decade.

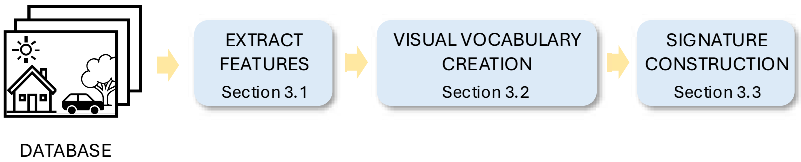

Figure 1: The visual bag of visual words framework: given a set of images (database), feature extraction is performed for all these images (Section 3.1). From all the features, the visual vocabulary is created (Section 3.2). Then, the signature of each image is built using the visual words that compose the vocabulary (Section 3.3).

3 SEMANTIC AND VISUAL BAG OF WORDS

In this work, a hierarchical bag of words is used to solve the localization problem in the navigation of a mobile robot. This hierarchical bag of words is composed of two levels. The higher level is based on semantic information, while the lower level is supported by visual information.

As it has been mentioned, the bag of visual words method involves creating a vocabulary which is composed of representative visual words that are the results of clustering the visual descriptors extracted in an image.

The bag of visual words technique consists of the extraction of visual features which are then clustered in order to create a set of visual words (vocabulary or codebook). After that, each image is represented by a signature that encodes the frequency with which each visual word appears in the image. These components are shown in Figure 1, and described in detail within the following sections.

3.1 Feature Extraction

In this work, the features used are the result of combining local feature descriptors (see Section 3.1.1) and semantic information (see Section 3.1.2). Both processes are carried out in parallel as shown in Figure 2.

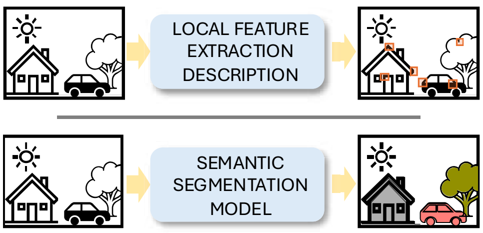

Figure 2: The upper part of the figure corresponds to the extraction of local features (points and its feature descriptors, Section 3.1.1). On the other hand, the bottom part of the figure corresponds to the extraction of semantic segmented map using a semantic segmentation model (Section 3.1.2).

3.1.1 Local Features

This stage is divided into two steps. Firstly, the distinctive local points are identified for each image in the database. This step is known as local feature extraction. These points can be corners, edges or blobs. Secondly, the extracted local points are represented by a feature descriptor that extracts visual features of its neighbourhood.

There are several techniques for this purpose, such as SIFT (Scale-Invariant Feature Transform) (Lowe, 2004), SURF (Speeded-Up Robust Features) (Bay et al., 2006) and ORB (Oriented FAST and Rotated BRIEF) (Rublee et al., 2011).

3.1.2 Semantic Segmentation

Semantic segmentation is a technique that aims to assign a semantic class to each pixel in the image.

Although the diagram presented in Figure 2 shows a block corresponding to the extraction of the semantic segmentation map using a semantic segmentation model, it is important to note that this step is not carried out in this work, as the semantic segmentation maps have been previously generated and are part of the dataset, as it will explained in Section 4.1.

3.1.3 Fusion of Semantic and Visual Information

A semantic segmentation map and a set of local points with their corresponding feature descriptors are available for each image. The goal is to have a semantic label associated to each local feature descriptor. For that end, given a local point, its coordinates are employed to extract the semantic label encoded in the map at this pixel. This label is then assigned to the feature descriptor of this local point.

At the end of this step, the visual descriptors have been classified into different semantic categories.

3.2 Visual Vocabulary Creation

In a bag of visual words algorithm, visual feature descriptors extracted from all images in the database are grouped into k clusters based on their similarity. This

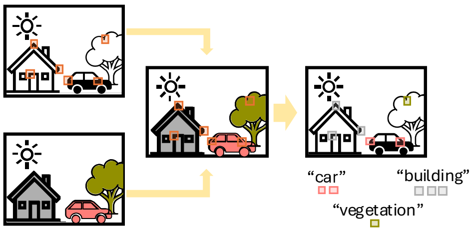

Figure 3: The local points detected are projected on the semantic segmented map to obtain its semantic label.

stage requires the use of a clustering algorithm such as k-means. This is followed by the creation of the visual vocabulary. The visual words are the centroids of the clusters and the size of the vocabulary is equal to the number of clusters (k).

In this work, no visual vocabulary is created for all extracted feature descriptors, but a single one is generated for each semantic category.

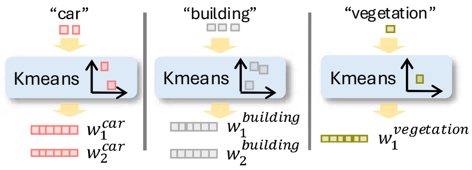

Figure 4: A bag of k visual words is created for each semantic class. Given a set of local feature descriptors classified as class i (e.g. car), a k-means algorithm is used to assemble these descriptors into k clusters (e.g. two) and to extract the centroids of each one. These centroids are the visual words of this class ($w_j^{class_i}$ j = 1, ..., k).

3.3 Signature Construction

In a bag of visual words algorithm, each image is represented by a one-dimensional vector with a length equal to the number of visual words (vocabulary size) where each element encodes the number of times each visual word appears in the image. In other words, the signature is a histogram of the frequencies of visual words.

In this work, there is not a unique bag of visual words, but a bag of visual words for each semantic category. Then, if the number of semantic categories to be considered is N, the signature is a set of N frequency histograms. It will be identified in the experimental section (Section 5) as BoSVW.

In addition, we propose replacing the frequency histogram with a vector in which each cell encodes the sum of the cosine distance between the local features and the visual word. It will be identified in the experimental section (Section 5) as BoSVW*.

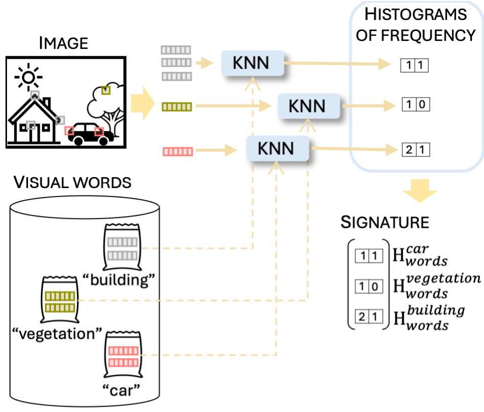

Figure 5: Once the feature descriptors have been divided into different groups based on the semantic information, a K nearest neighbour algorithm is used to find the most similar visual word with the same semantic class. After obtaining a histogram for each semantic group, the signature is constructed.

4 EXPERIMENTAL SETUP

In this section, the visual localization problem is solved using several configurations of bag of visual words method. The main purpose is to evaluate these configurations and select the one that provides the best results. It is important to note that the full content-based image retrieval algorithm proposed by Ouni et al. (2022) has not been implemented. Only the signature construction module explained in Section 3.1 of (Ouni et al., 2022) was implemented in our framework. It will be identified in the experimental section (Section 5) as BoSW. Table 1 shows a brief definition of each of these configurations, their IDs used in Section 5 and a brief description of the image signature.

The images used are equirectangular (more information on the dataset is given in Section 4.1) in order to evaluate these methods also for this type of images, since spherical images present some benefits in mobile robot navigation (such as their wide field of view).

The algorithm executed in all four cases is the same except for the encoding image part, where the procedure is different for each type of signature (or distance for BoSVW). In all four cases, the measure chosen to compare the signatures of the query image and those in the database is the cosine distance. Since these are two-dimensional image signatures in all four cases, the rows are concatenated to obtain a one-dimensional vector. After that, the vector is normalized using L1. The cosine distance is then calculated, as indicated, in order to obtain the most similar

Table 1: The different configurations of the bag of words method that are evaluated. The first column shows the ID used for each configuration when displaying the results, while the third column describes the signature of each.

ID	Method	Signature
BoSW	Bag of Semantic Words (Ouni et al., 2022)	MxN matrix where the width N corresponds to the size of the visual descriptor and the height M corresponds to the number of semantic classes.
BoSVW	Bag of Semantic Visual Words	A set of M (number of semantic classes) one-dimensional frequency histograms with a length equal to K.
BoSVW*	Bag of Semantic Visual Words (no frequency).	A set of M (number of semantic classes) one-dimensional vectors with a length equal to K where each bin is the sum of the cosine distances instead of the count.

image (retrieved image) in the database based on this.

All signatures have in common that they use visual information. In this work, ORB (Rublee et al., 2011) has been chosen to obtain the local points and their corresponding feature descriptors.

In the cases of BoSVW and BoSVW*, the vocabulary size (k) is initially fixed for all semantic classes and its value is 10. If any semantic class has a number of feature descriptors less than 10, then the vocabulary size of this class is equal to the number of feature descriptors.

4.1 Dataset

The dataset employed in this paper is KITTI-360 (Liao et al., 2023). For the image collection, the authors equipped a station wagon with two fisheye camera positioned to each side. Both fisheye cameras have 180 degrees of field of view so that a full view of the scene is captured.

Before carrying out the experiments, each pair of fisheye images was converted into a single equirectangular image. For that, the calibration provided in the dataset is used to convert each fisheye image into equirectangular (fisheye image projection to unit sphere). Then, a polynomial transformation proposed by Flores et al. (2024) is used to align both equirectangular images. Figure 6 shows an example (equirectangular image) after performing this process.

Figure 6: An equirectangular image generated from a pair of fisheye images of the KITTI-360 dataset (Liao et al., 2023).

The dataset of images has been divided into two subsets: database and query. For the first subset, an image was selected every 10 meters of the trajectory, taking the first captured image as the starting image. Thus, the database consists of a total of 791 images. The images of the dataset not selected have been considered as query images. This means that 9723 images constitute the query set.

As mentioned above, the semantic segmentation maps were not obtained during the running process (obtained beforehand). The semantic segmentation model employed for this purpose is SegFormer (Xie et al., 2021). A semantic segmented map can be visualized in Figure 7. Due to the fact that it is a 380 vision system, part of the station wagon appears in the image. However, it is not part of the scene so it has been labelled as unlabelled (black pixels in Figure 7), and this class has not been taken into account for the image signature construction.

Figure 7: The semantic segmentation map of the image shown in Figure 6. It was generated using SegFormer (Xie et al., 2021).

4.2 Evaluation Protocol

4.2.1 Distance Difference

The distances between the query image (q) and the retrieved image (r) will be analysed. Then, given a query image captured at position XYZ_q, the image retrieval algorithm is executed, which returns the database image most similar to the query image (i.e. retrieved image). The retrieved image was acquired at XYZ_r. Both positions (XYZ_q and XYZ_r) are extracted from the pose file provided by the dataset. The distance units are meters and are calculated as follows:

$$dist_{q-r} = \sqrt{(X_q - X_r)^2 + (Y_q - Y_r)^2 + (Z_q - Z_r)^2} \quad (1)$$

4.2.2 Average Recall (AR) at 1

For each query image, a retrieved image is recovered from the database after applying the image retrieval method. Since the dataset provides the poses in which all images were captured, after acquiring the retrieved image, the distance between the pose of the retrieved image and the query image ($dist_{q-r}$) is calculated using equation (1).

If the distance is lower than 20 meters, the recall for this query image (I_{q_i}) is one. Otherwise, the recall will be zero.

$$R@1_{q_i} = \begin{cases} 1 & \text{if } dist_{q-r} < 20\text{meters} \\ 0 & \text{if } dist_{q-r} \geq 20\text{meters} \end{cases} \quad (2)$$

The evaluation measure is the average value of all recall values after executing the method for all query images (n images):

$$AR@1(\%) = \frac{\sum_{i=1}^{n} R@1_{q_i}}{n} \cdot 100 \quad (3)$$

5 RESULTS AND ANALYSIS

5.1 BoSVW: Feature Extraction in the Database

To create the vocabulary, it is important to extract features. In this section, we analyse the visual features extracted from the images in the database, specifically how many are associated with each of the semantic classes. This can be observed in Figure 8 by means of a graph chart, where the height of each bar represent the number of visual features that are classified for each semantic class.

As it can be seen, the semantic class with the highest number of features is "vegetation", with a total of 374321. In contrast, "train" is the class with the lowest number, a total of 4. Other semantic classes with a high number of associated local features are "building", "car" and "sky", in that order.

5.2 Comparison and Evaluation

This section compares the different image signatures presented in Table 1 using the evaluation protocols described in Section 4.2. For the case of BoSVW, two experiments have been carried out. The first experiment used the Euclidean distance to create the histogram (i.e. during the K Nearest Neighbour process).

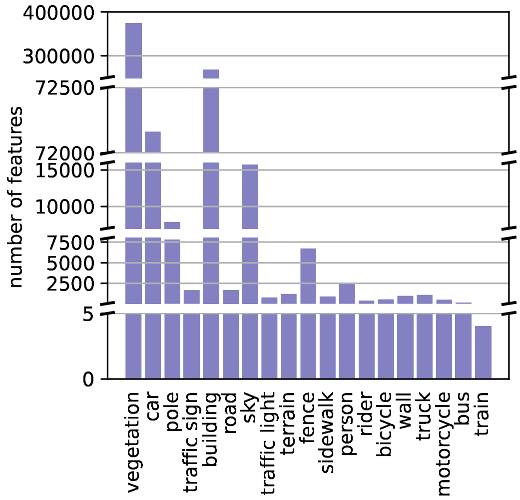

Figure 8: Number of ORB local features (y-axis) for each semantic class (x-axis).

The second experiment uses the cosine distance. The aim is to determine whether it has an influence on the results.

5.2.1 Evaluation in Terms of Average Recall at 1

First, the different bag of words methods mentioned above (see Table 1) are evaluated in terms of Average Recall at 1 (AR@1). The results are shown by means of bar graphs which can be observed in Figure 9. Each bar indicates the AR@1 achieved for each signature type.

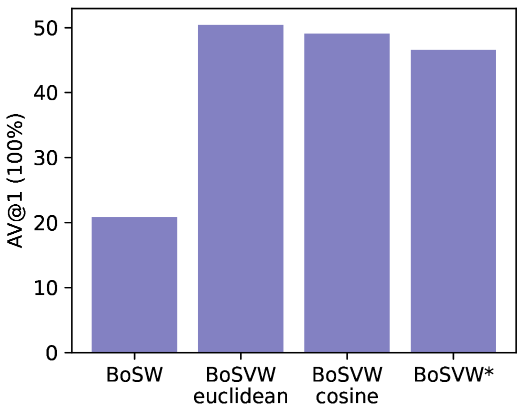

Figure 9: Results of Average Recall at 1 (AR@1) calculated after using the four image signatures evaluated.

With regard to the results shown in this figure, the value of AR@1 for the BoSW method is equal to 20.806, for BoSVW using the Euclidean distance it is equal to 50.417, meanwhile using the cosine distance the value is 49.069, and for BoSW* it is equal

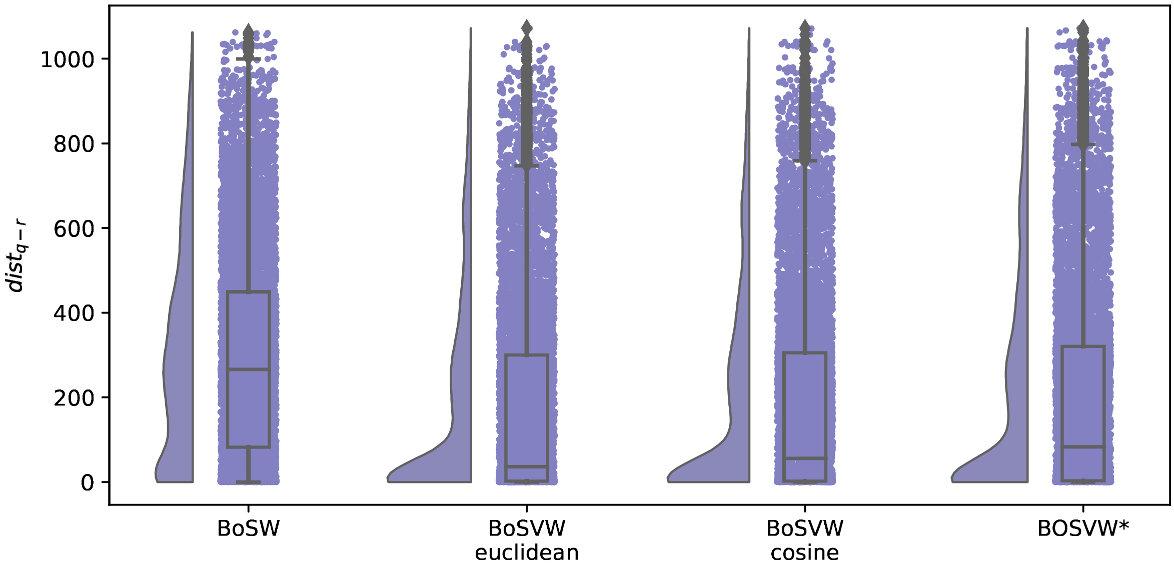

Figure 10: The distance between the position of the query image and the retrieved image using the different signatures.

to 46.529. As can be seen, a higher value of recall in place recognition is achieved for BoSVW variations than for BoSW. The results show that using frequency (BoSVW) instead of distance (BoSVW*) achieves better results. In terms of the distance to find the visual words, the Euclidean distance provides better results than the cosine distance.

In summary, we propose two image signatures (BoSVW and BoSW*) in which each semantic category has its own associated bag of visual words from which its descriptor will be obtained, rather than the descriptor being a single visual descriptor (BoSW). The main objective of this paper is to solve the localization problem by means of place recognition. According to this initial evaluation, the image signatures we propose increase the AR@1 value when they are used to address image retrieval in a set of images from the same environment (a trajectory followed by a mobile robot). In other words, the proposed image signatures improve the capability of the algorithm to return an image of the database with a distance of less than 20 metres from the query image on more times.

5.2.2 Evaluation in Terms of Distance Difference

Second, the different methods are also evaluated in terms of distance difference in meters. For this purpose, the distance in meters between each query image in the query set (9723 images) and its retrieved image is calculated using (1). For each type of signature, all of these 9723 distances are displayed using rain cloud plots, which help to analyse the distribution of distances as probability density, and at the same time, key summary statistics (such as median and quartiles) can be visualized. This graph can be seen in Figure 10. It is important to note that the database set is made up of an image every 10 meters on the trajectory (see Section 4.1). Then, the range of Euclidean position distances between the query images and the nearest database images (ground truth) is between 0 and 5 meters. Therefore, the committed localization error is considerable in the cases (points in the graph) where the distance between the query image and the retrieved image ($dist_{q-r}$) is much greater than 5 meters.

At first view, we can see that, for all four types of signatures, the highest concentration of points is found on distances below 100 metres approximately. In addition to this, it is clear that the three types of BoSVW have a higher peak than BoSW in this interval. As for the box plots, BoSW has a higher median value (around 265 meters) than the other three image signatures. Also, the lower whisker of the BoSW boxplot has a longer distance compared to the others. Focusing on BoVW types, the results are better when using the frequency histogram, the median value is around 36 metres when the distance is Euclidean and around 55 metres when it is cosine. However, the median value is around 84 metres when using the vector representing the sum of the cosine distance to the visual words.

In this section, the image signatures have not been evaluated only on the condition of finding the most similar one within a ratio, as in the previous evaluation, but rather all distances are shown after running the image retrieval algorithm for all query images. It for each signature. Taking these results into account, the proposed signatures most frequently return as the most similar image from the database that is in a closer position to the query image, achieving a more refined localization.

6 CONCLUSIONS

The main objective of this work is to solve the task of place recognition for a mobile robot navigating in a known outdoor environment. The method used for this is image retrieval using equirectangular images as input. Image retrieval relies on the fact that images are represented in a way that their significant features are described (image signature).

In relation to this, three types of image signatures are evaluated and compared in this work for place recognition, when a mobile robot navigates a trajectory in a previously visited environment. All the implemented signatures combine semantic and visual information. The first one (i.e. BoSW) was proposed by Ouni et al. (2022) whereas the other two variations are proposed in this paper (i.e. BoSVW and BoSVW*). The BoSW image signature is a matrix in which each row is the centroid of a set of visual feature descriptors belonging to the same semantic class. The number of rows is equal to the number of semantic classes and the number of columns is equal to the size of the local visual feature. In the case of BoSVW the rows are histogram of frequency of visual words.

After the experiments, the results in terms of recall at one determine that BoSVW using the Euclidean distance during the image encoding step provides the highest value. In contrast, the lowest recall value is achieved using BoSW. Apart from this evaluation measure, the distances between the position of the query image and the position of the image retrieved by the method using each image signature is also analysed. The use of a Euclidean distance achieves a lower distance in more times than the cosine distance for BoSVW.

Therefore, it can be concluded that creating a bag of visual words for each semantic category, such as proposed in this paper, rather than a single visual descriptor, improves the results on the problem of place recognition. Additionally, if each category is represented by a frequency histogram, the localization is more accurate than using a vector that encodes distances.

In summary, the evaluations show that the implementation of the proposed signatures in an image retrieval algorithm for place recognition provides better results.

In this work, only image signatures that merge semantic and visual information have been evaluated and compared to solve the place recognition. Taking it into account, we propose as a future work to extend this comparative evaluation to other algorithms (such as these that use only visual information). In the same line, other possible future work can be study these signatures using other local features, both using traditional extraction methods and Deep Learning methods. Finally, other future work could be to research whether the proposed signatures can be improved by finding the optimal value of clusters for each vocabulary size in each category, rather than this parameter being fixed for all semantic categories as it is in this work.

ACKNOWLEDGEMENTS

This research work is part of a project funded by ”AYUDAS A LA INVESTIGACIÓN 2025 DEL VICERRECTORADO DE INVESTIGACIÓN Y TRANSFERENCIA” of the Miguel Hernández University and 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 and part of the project CIAICO/2023/193 funded by Generalitat Valenciana.

REFERENCES

Alfaro, M., Cabrera, J., Jiménez, L., Reinoso, O., and Payá, L. (2024). Triplet Neural Networks for the Visual Localization of Mobile Robots:. In *Proceedings of the 21st International Conference on Informatics in Control, Automation and Robotics*, pages 125–132, Porto, Portugal. SCITEPRESS - Science and Technology Publications.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). SURF: Speeded Up Robust Features. In Leonardis, A., Bischof, H., and Pinz, A., editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg. Springer.

Cabrera, J. J., Santo, A., Gil, A., Viegas, C., and Payá, L. (2024). MinkUNeXt: Point Cloud-based Large-scale Place Recognition using 3D Sparse Convolutions. arXiv:2403.07593.

Dubey, S. R. (2022). A Decade Survey of Content Based Image Retrieval Using Deep Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2687–2704.

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

Forcen, J. I., Pagola, M., Barrenechea, E., and Bustince, H. (2020). *Co-occurrence* of deep convolutional features for image search. *Image and Vision Computing*, 97:103909.

Li, X., Yang, J., and Ma, J. (2021). Recent developments of content-based image retrieval (CBIR). *Neurocomputing*, 452:675–689.

Liao, Y., Xie, J., and Geiger, A. (2023). 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.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Mansoori, N. S., Nejati, M., Razzaghi, P., and Samavi, S. (2013). Bag of visual words approach for image retrieval using color information. In *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, pages 1–6. ISSN: 2164-7054.

Ouni, A., Chateau, T., Royer, E., Chevaldonné, M., and Dhome, M. (2023). An efficient ir approach based semantic segmentation. *Multimedia Tools and Applications*, 82(7):10145–10163.

Ouni, A., Royer, E., Chevaldonné, M., and Dhome, M. (2022). Leveraging semantic segmentation for hybrid image retrieval methods. *Neural Computing and Applications*, 34(24):21519–21537.

Rani, S., Kasana, G., and Batra, S. (2025). An efficient content based image retrieval framework using separable CNNs. *Cluster Computing*, 28(1):56.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571. ISSN: 2380-7504.

Srivastava, D., Singh, S. S., Rajitha, B., Verma, M., Kaur, M., and Lee, H.-N. (2023). Content-Based Image Retrieval: A Survey on Local and Global Features Selection, Extraction, Representation, and Evaluation Parameters. *IEEE Access*, 11:95410–95431.

Vilella-Cantos, J., Cabrera, J. J., Payá, L., Ballesta, M., and Valiente, D. (2025). MinkUNeXt-SI: Improving point cloud-based place recognition including spherical coordinates and LiDAR intensity. arXiv:2505.17591.

Wang, T.-H., Huang, H.-J., Lin, J.-T., Hu, C.-W., Zeng, K.-H., and Sun, M. (2018). Omnidirectional CNN for Visual Place Recognition and Navigation. arXiv:1803.04228.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090.

Xu, S., Chou, W., and Dong, H. (2019). A Robust Indoor Localization System Integrating Visual Localization Aided by CNN-Based Image Retrieval with Monte Carlo Localization. *Sensors*, 19(2):249.

Zeng, Z., Zhang, J., Wang, X., Chen, Y., and Zhu, C. (2018). Place Recognition: An Overview of Vision Perspective. *Applied Sciences*, 8(11):2257.