الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Département : Génie électrique Et Informatique Industrielle

Mémoire de fin d'étude en vue de l'obtention du Diplôme de
Master

Spécialité
Automatique et Informatique Industrielle

- Thème -

Place Recognition for Mobile Robots

Réalisé par
LAHOUAL Yassine

Les membres de Jury:

M. Abbadi belkacem	Président
Mme. Rebai karima	Promotrice
Mme. Bouchama samira	Examinatrice
Mme. Chouaf seloua	Examinatrice

Alger, le 15/09/2022
19 صفر من 1444 هـ

Année universitaire 2021 –2022

Place Recognition For Mobile Robots

LAHOUAL Yassine
Electrical Engineering and Computer Science
National Higher School of Technology
Algiers, Algeria
Y_lahoual@enst.dz

REBAI Karima
Electrical Engineering and Computer Science
National Higher School of Technology
Algiers, Algeria
karima.rebai@enst.dz

***Abstract*— The ability of autonomous vehicles to navigate in challenging maps with continuously changing conditions is so impressive and still improving. A key component that enables these intelligent vehicles to overcome such conditions and become more autonomous is the precision of the perception and localization systems. As part of the localization system, place recognition has benefited from recent developments, especially with the emergence of deep learning frameworks. This article studies the current state-of-the-art approaches and methods used in place recognition based on deep learning. Firstly, introduces the full picture of place recognition concepts. Then, addresses its challenges. Finally, how deep learning can help solving those problems. The DL-based place recognition can be categorized into supervised, unsupervised, semi-supervised categories. Therefore, this article will dive into each category including the advantages and disadvantages of each in place recognition, without forgetting to mention some of the related works.**

Keywords*— *Autonomous Vehicles, Place Recognition, Deep Learning.

I. INTRODUCTION

In the last decade, a lot of studies and improvements have been made in autonomous vehicles and robotic vision, because of that, intelligent vehicles are becoming more capable of navigating in complex and wide environment. For a better autonomous operation, a strong navigation system is needed. The key to achieve a higher level of autonomy is related to the adopted solution for Simultaneous Localization and Mapping (SLAM). The purpose of SLAM is that the vehicle needs to build the environment's map and localize itself within its surroundings then correct mis-localization in the unknown environment. Visual Place Recognition (VPR) is the heart component of SLAM and its often named as closed loop detection; a system which can decide whether the place has previously been visited or not.

Thanks to deep convolutional neural networks which have a strong generalization power to describe the places with robustness to conditional and viewpoint variations which are the main challenges for any VPR system to beat. The influence of DL frameworks in place recognition is, in particular, observable when considering the vast amount of place recognition works published in the last few years that resort to such methods. Therefore, the contribution of this work is to provide a comprehensive review of the recent methods and approaches.

The remainder of this paper is organized as follows. The next Section II will give a definition of recognition for mobile robot localization which will include simultaneous localization and mapping (SLAM) and visual place recognition (VPR) systems, followed up with major challenges that VPR can face in Section III. Then, Section IV will talk about deep learning-based approaches for visual place recognition and into each category which are supervised, unsupervised and semi-supervised. Finally the conclusion in Section V.

II. PLACE RECOGNITION FOR MOBILE ROBOT LOCALIZATION

A. Simultaneous Localization and Mapping (SLAM)

To achieve robotic localization, the system requires a map to recognize a previously visited location. Simultaneous Localization and Mapping is the process of creating such maps of the surrounding environment for the purpose of localization (SLAM). SLAM has been a hot topic in the robotics and computer vision sectors for the past couple decades. It is made up of the following elements:

Mapping: It's the process of internally producing and storing a representation of the outside world. The sort of environment and sensory information used to create world maps might differ. As seen in "Fig. 1", occupancy grid maps divide the world into uniformly spaced discrete sections or cells. Occupancy grid maps presume that each cell is occupied or unoccupied. For occupancy grid maps, few assumptions are made; each cell is either free or occupied, and each cell is independent of the others. Each region has a probability to be occupied by an obstacle probability based on a cumulative knowledge and statuses. A 0.5 probability to be an obstacle have been assigned to those regions that are neither occupied nor explored. Such occupancy-based grid maps are ideal for crowded indoor situations since they can be easily created using laser or sonar-based data but demand a lot of memory. These maps are commonly used in obstacle avoidance algorithms. Feature maps, on the other hand, are compact, sparse, and consist of discrete landmarks with their coordinates. With information about landmarks collected from the vision sensors, they are ideal for use in outside contexts. Pose graphs are a type of map that depicts the robot's trajectory graphically. The nodes orientation describes the robot's pose (position and orientation). Edges describe the spatial connection between the poses, can be used for loop closure.

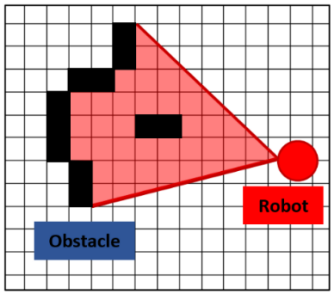


Fig. 1 An occupancy grid map, light-red is the vision of the robot, black is the obstacle.

Localization: The method of tracking the position of a robot inside a map using different types of sensors is known as localization. This module takes the acquired data by sensors data and generates a belief about where robot is on the map. Generally, visual sensors are the most used sensors known by the rich provided information. The belief is used by the algorithm to determine if the current location has been viewed before or is new. When two places look alike, it's typically because they're in the same area. This assumption, however, does not hold true when there are substantial weather and lighting changes that affect the visual measurements, which could cause the robot to incorrectly localize within the map.

B. Visual Place Recognition (VPR)

Visual place recognition is one of the most important parts of visual Autonomous Vehicles, and can be defined as the task of identifying the same locations visited despite significant changes in appearance and viewpoint at different periods of time. Visual place identification is an important part of Spatial Artificial Intelligence, as it allows robotic platforms and intelligent augmentation platforms like augmented reality devices to sense, recognize and interpret the physical environment. It has gained lot of attraction in recent times [1] due to its many advantages over other sensor modalities such as low cost, low power consumption, easy setup and ability to work even in areas with no Global Positioning System signal (In an outdoor, GPS offers a cheap solution for localization). Also, the potential applications of place recognition in the field of loop closures, localization, etc., making it a critical challenge to tackle under the umbrella of Simultaneous Localization and Mapping (SLAM). Place recognition, in particular, is a reliable way for removing accumulated mistakes in SLAM because revisiting sites allow loop-closure in built maps. Therefore, the task of VPR is to find associations between one or multiple query images with a database of images of known positions (places), known as image retrieval task, see "Fig. 2". So, it can detect loop-closure during robot's motion and contribute in the improvement of SLAM quality. Due to its importance for long-term operation, a particularly active research field is recognizing places despite severe changes of the environmental conditions induced by day-night cycles, changing weather, or seasons. At its core, VPR is an image retrieval problem and many developments and results from the computer vision community significantly contributed to the state of the art. However, the application in the context of mobile robotics frames the image retrieval task and typically provides additional structure and information.

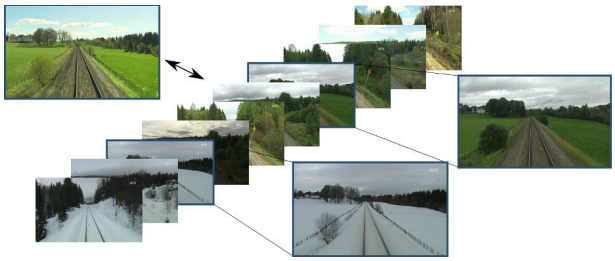


Fig. 2 Running VPR system and comparing the most similar between query images and the input image up on the left corner.

It is an active field of research and many different approaches have been proposed and evaluated in many different experiments [2].

III. Major Challenges for VPR

The main goal of visual place recognition is to match two images so that the system can determine if it has seen that location previously. And it is notably that place recognition approaches are becoming more and more sophisticated as the environment and operation conditions become more similar to real-world situations. The present state-of-the-art in place recognition techniques, for example, can operate over large regions in real-world conditions situations with unique performances. It has been highlighted as important issue to be able to locate in and create maps of dynamic environments [2]. Despite these achievements, place recognition task poses a number of challenges due to three major reasons. This section discusses over these challenges which they occur when:

- two distinct places have similar appearances, and the same place with appearance changes over time due to light conditions like day-night variation, weather, seasonal or structural changes as shown in "Fig. 3".
- same places are observed from different positions and/or with different viewpoints.
- scalability of the proposed approaches in large environments.

These problems must be overcome, to allow effective robot's localization and robust place recognition and as a result a perfect long-term localization.

A. Appearance Change and Perceptual Aliasing

Many studies have focused on appearance-changing and perceptual aliasing in particular. Because autonomous robots move for long periods of time, their perception systems must deal with changing that may happen over time, such as changing weather or seasonal conditions, or structural modifications. While the appearance change: issue occurs when the appearance of the same place varies over time, the perceptual aliasing occurs when distinct places appear to be the same. These factors have an important impact on place recognition performance and consequently on loop closure detection.

Various research works have addressed these issues. The belief generation perspective and Sequence-based matching techniques like [3] are emphasized as extremely successful in these conditions. The task of aligning a pair of template and query sequences is known as sequence matching, and it may be done via minimal cost flow, or probabilistically using Hidden Markov models. Place modeling approaches can also handle this problem: extracting condition-invariant features [4] , extracting for example features using convolutional neural network and other architectures of Deep learning [5], Descriptor normalization [6].



Fig. 3 Illustration of the seasonal environment changes. Images taken from the Oxford Robotcar, and Norland dataset.

B. Viewpoint Changing

Place recognition is especially difficult when revisiting a location from another [6]. This is especially true for methods that rely on sensors with a limited field-of-view or lack geometrical sensing capabilities. These sensors only catch a fraction of the environment while visiting a location, and upon revisiting from a different angle or position, the scene may seem differently or even more features may be felt, resulting in an entirely different place model.

In some studies, the use of panoramic cameras or 3DLiDAR [7] have been proposed and tested, this reduces the effect of how future visitors interpret orienting places. As a result, using sensors and methods that aren't affected by orientation (also known as viewpoint-invariant) makes location recognition more reliable and robust. In other works, visual-based techniques have turned to semantic-based features to solve these issues [8].



Fig. 4 Example of two images of the same place observed from different positions.

C. Scalability

Scalability is another important aspect of place recognition. As self-driving cars operate in wider regions, more locations are visited and maps grow in size, increasing processing load and lowering inference efficiency. Therefore, some approaches like: efficient indexing [9], scalar quantization [10], Hidden Markov Models (HMMs) [11], or learning regularly repeating visual patterns [12] are all methods for increasing inference efficiency.

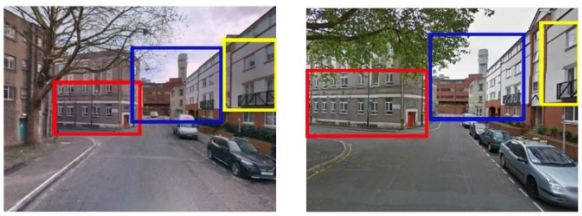


Fig. 5 Example of two images of the same place observed from different distance.

IV. Deep Learning-Based Approaches For VPR

Autonomous vehicles (AV) are improving their ability to navigate in complicated and changing conditions of the environment. The greatest development of the vision and localization systems is a key component that enables these intelligent cars to overcome such situations that are mentioned in section III and become more autonomous. To this end, place recognition has benefited from recent breakthroughs in other perceptual tasks such as place classification and object recognition , notably the rise of deep learning (DL) frameworks, as part of the localization system.

Deep learning's success in the computer vision sector has inspired a burst of early research on its capabilities for visual place recognition. This section study latest deep learning-based place recognition approaches that can be divided into supervised, unsupervised, semi-supervised "Fig. 6".

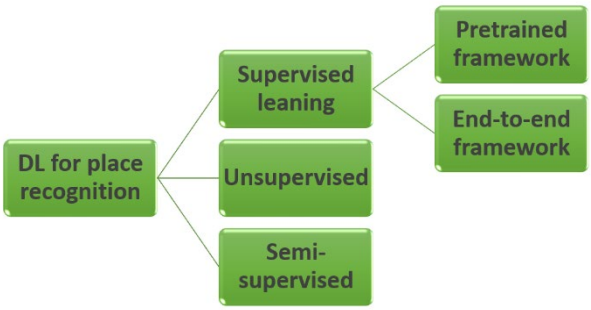


Fig. 6 Organization of recent DL-based place recognition approaches.

A. Supervised place recognition

Firstly, place recognition approaches by using supervised deep learning consist on using labeled data. Supervised machine learning develops a function that transfers an input representation (such as images) to an output representation (such as a descriptor). This function takes the form of weights in a network with several layers in deep learning. The weights are learnt in stages, with the first step calculating the difference between predictions and ground truth (real desired output), and the second step backpropagating the error using gradient descent [13]. This method (i.e., assessing errors and modifying weights) is repeated until the network's predictions are accurate. The capacity to automatically learn features from the training data is a benefit of such a learning process, particularly when utilizing convolutional neural networks (CNN).

Deep supervised learning provided advancements in place recognition. The capacity of CNNs to extract features in particular resulted in more descriptive place models, which improved place matching. Early techniques relied mostly on pre-trained CNNs that had already been trained on other visual tasks. However, new techniques have recently enabled training of DL networks on place recognition tasks named end-to-end frameworks [14].

1) Pre-trained-based Frameworks

Pre-trained place recognition frameworks are techniques that extract features using pre-trained CNN models that were previously trained on other perceptual tasks such as object recognition, place categorization [6] and segmentation [15].

There are three types of works that use such models: holistic-based, landmark-based, and region-based. The first is Holistic-based techniques relate to works that employ all activations from a layer as a descriptor and feed the entire image to a CNN. Because CNNs are hierarchical, the various layers contain features with different semantic interpretations. In contrast to holistic-based techniques, which convert all picture features to descriptors, the second landmark-based approaches convert just the features from the recognized landmarks to descriptors like Edges. And the third is region-

based approaches which focus on local features as the landmark-based technique do, however, region-based approaches send the full picture to the DL model and use just the salient activation in the CNN layers for descriptors. "Fig. 7" illustrates such approaches applied to an input image.

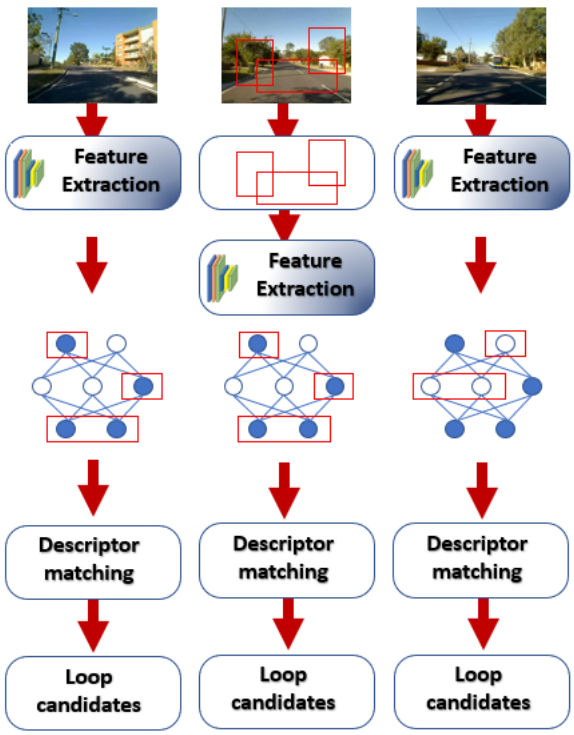


Fig. 7 Pre-trained frameworks: left holistic-based, middle landmark-based and right region-based.

2) End-to-End Framework

End-to-end frameworks, in contrast to pre-trained frameworks, use machine learning algorithms to learn the feature representation and get a descriptor directly from sensor data while training on a place recognition problem. The defining of the training objective: that is, what are the networks optimized for and how are they optimized, is an important part of end-to-end learning. Networks are often optimized in place recognition to provide unique descriptors that can identify the same physical location independent of appearance or viewpoint.

The loss function is especially important during the training phase, since it reflects the mathematical interpretation of the training objective and hence determines the optimization process's successful convergence. An example in [16] using quadruplet loss function to train networks for place recognition.

To deal with the problem of place recognition using end-to-end approaches, some works proposed the use of camera, Lidar [7] and radar [14]. The developed approaches are directly related to the nature of data. The beneficial from low-cost system that translates camera views to driving control inputs is possible with a camera-based end-to-end driving neural network. These networks are appealing because they can extract image features without extracting them using convolutional methods called hand-crafted features. Yin et al. [17] propose a place recognition and metric pose estimation strategy using 3D point clouds to handle the global localization challenge.

B. Unsupervised Place Recognition

The supervised learning techniques mentioned above provide great results when learning discriminative location models. However, like with supervised DL-based methods, these methods have the difficulty of requiring a large quantity of labeled data to function successfully. Unsupervised learning, unlike supervised learning, does not require labeled data, which is beneficial when annotated data is unavailable or sparse.

For domain translation, unsupervised techniques such as Generative Adversarial Networks (GAN) are used [18]. Latif et al. [18] describe one such technique, which treats the cross-season place recognition issues as a domain translation job. GANs are used to learn the relationship between two domains without requiring cross-domain image correspondences. Two connected GANs are provided as the suggested architecture. The discriminator includes an encoder network followed by two completely linked layers, whereas the generator includes an encoder-decoder network. The discriminator's output is utilized as a description in place recognition.

Merril and Huang [19] present an autoencoder-based strategy to manage feature embedding for visual loop closure, in contrast to previous efforts that were mostly focused on GAN approaches. This unsupervised technique is especially built to map pictures to a Histogram of Oriented Gradients (HOG) descriptor space, rather than rebuilding original images. The autoencoder network is trained with a pre-processing stage as input, which employs two basic geometric vision techniques: histogram of oriented gradients (HOG) and projective transformation (homography). HOG allows for compression while keeping important details. The projective transformation, on the other hand, enables the connection of pictures from different viewpoints. The network's design is simple, allowing for real-time close-loop detection with no dimensionality reduction.

C. Semi-Supervised Place Recognition

Semi-supervised technique is a mixed between of supervised and unsupervised methods. For the cross-domain challenge, the combination of these two learning methodologies is especially useful. Instead of translating from one domain to another, these learning approaches are used to learn traits that are independent of the look of the domain.

Wang et al. [20] offer a method that combines weakly supervised and unsupervised to learn domain-invariant features for cross-domain visual place recognition. Three key elements make up the proposed architecture: an attention module, an attention-aware VLAD module, and a domain adaptation module. The unsupervised branch uses a multi-kernel maximum mean discrepancy also known as MK-MMD loss function, whereas the supervised branch uses a triplet ranking loss function.

Tang et al. [21] present a self-supervised learning strategy to distinguish between place-related and domain-related characteristics. The proposed approach's architecture is a modified autoencoder for adversarial learning. The disentanglement of the two feature domains is accomplished via adversarial learning, which limits the learning of domain specific features.

For localization and mapping, Dube et al. [22] suggest SegMap, a data-driven learning technique. The method learns object segments from 3D point clouds using an autoencoder-like architecture as the basic framework. The framework is utilized for two tasks: classification (supervised) and reconstruction (unsupervised).

V. Conclusion And Discussion

This article presents the recent state-of-art methods applied on place recognition approaches, emphasizing the recent developments on deep learning frameworks, namely supervised, unsupervised, semi-supervised.

Each of these frameworks was described briefly. In supervised approaches labeled data is needed, the pre-trained frameworks tend to resort to semantic information by detecting landmarks or leveraging regional activation from Convolutional Neural Networks layers. On the other hand, among the end-to-end frameworks, the purpose is training the model directly on place recognition using sensory data from the camera, 3D LiDAR, or RADAR. Unsupervised approaches do not require labeled data, the main application of unsupervised approaches, such as GANs and autoencoders, is to address the domain translation problem. While in semi-supervised, which is a combination of the two supervised and unsupervised methods, the works address the cross-domain problem, however instead of translating a source domain into a target domain, these works seek to obtain a descriptor space that is invariant to domains. Some recently related works were highlighted all of them seek to increase performance of visual place recognition system.

References

[1] T. Barros, R. Pereira, L. Garrote, C. Premebida, and U. J. Nunes, "Place recognition survey: An update on deep learning approaches," pp. 1–18, 2021, [Online]. Available: http://arxiv.org/abs/2106.10458

[2] S. Lowry *et al.*, "Visual Place Recognition: A Survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, 2016, doi: 10.1109/TRO.2015.2496823.

[3] O. Vysotska and C. Stachniss, "Effective visual place recognition using multi-sequence maps," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1730–1736, 2019, doi: 10.1109/LRA.2019.2897160.

[4] P. Yin *et al.*, "A multi-domain feature learning method for visual place recognition," *Proc. - IEEE Int. Conf. Robot. Autom.*, vol. 2019-May, pp. 319–324, 2019, doi: 10.1109/ICRA.2019.8793752.

[5] L. Wu and Y. Wu, "Deep Supervised Hashing with Similar Hierarchy for Place Recognition," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 3781–3786, 2019, doi: 10.1109/IROS40897.2019.8968599.

[6] S. Garg, N. Suenderhauf, and M. Milford, "Don't look back: Robustifying place categorization for viewpoint- and condition-invariant place recognition," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 3645–3652, 2018, doi: 10.1109/ICRA.2018.8461051.

[7] L. Schaupp, M. Burki, R. Dube, R. Siegwart, and C. Cadena, "OREOS: Oriented Recognition of 3D Point Clouds in Outdoor Scenarios," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 3255–3261, 2019, doi: 10.1109/IROS40897.2019.8968094.

[8] S. Garg, N. Suenderhauf, and M. Milford, "Semantic–geometric visual place recognition: a new perspective for reconciling opposing views," *Int. J. Rob. Res.*, 2019, doi: 10.1177/0278364919839761.

[9] T. Cieslewski and D. Scaramuzza, "Efficient decentralized visual place recognition using a distributed inverted index," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 640–647, 2017, doi: 10.1109/LRA.2017.2650153.

[10] S. Garg and M. Milford, "Fast, Compact and Highly Scalable Visual Place Recognition through Sequence-based Matching of Overloaded Representations," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 3341–3348, 2020, doi: 10.1109/ICRA40945.2020.9196827.

[11] Y. Latif, A. D. Doan, T. J. Chin, and I. Reid, "SPRINT: Subgraph Place Recognition for INtelligent Transportation," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 5408–5414, 2020, doi: 10.1109/ICRA40945.2020.9196522.

[12] L. Yu, A. Jacobson, and M. Milford, "Rhythmic representations: Learning periodic patterns for scalable place recognition at a sublinear storage cost," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 811–818, 2018, doi: 10.1109/LRA.2018.2792144.

[13] A. Kanezaki, "Unsupervised image segmentation by backpropagation," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 1543–1547, 2018, doi: 10.1109/ICASSP.2018.8462533.

[14] S. Saftescu, M. Gadd, D. De Martini, D. Barnes, and P. Newman, "Kidnapped Radar: Topological Radar Localisation using Rotationally-Invariant Metric Learning," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 4358–4364, 2020, doi: 10.1109/ICRA40945.2020.9196682.

[15] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 2614–2620, 2017, doi: 10.1109/ICRA.2017.7989305.

[16] Z. Liu *et al.*, "LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, no. Iccv, pp. 2831–2840, 2019, doi: 10.1109/ICCV.2019.00292.

[17] H. Yin, L. Tang, X. DIng, Y. Wang, and R. Xiong, "LocNet: Global Localization in 3D Point Clouds for Mobile Vehicles," *IEEE Intell. Veh. Symp. Proc.*, vol. 2018-June, no. Iv, pp. 728–733, 2018, doi: 10.1109/IVS.2018.8500682.

[18] Y. Latif, R. Garg, M. Milford, and I. Reid, "Addressing challenging place recognition tasks using generative adversarial networks," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 2349–2355, 2018, doi: 10.1109/ICRA.2018.8461081.

[19] N. Merrill and G. Huang, "Lightweight Unsupervised Deep Loop Closure," *Robot. Sci. Syst.*, 2018, doi: 10.15607/RSS.2018.XIV.032.

[20] Z. Wang, J. Li, S. Khademi, and J. Van Gemert, "Attention-Aware Age-Agnostic Visual Place Recognition".

[21] L. Tang, Y. Wang, Q. Luo, X. Ding, and R. Xiong, "Adversarial Feature Disentanglement for Place Recognition Across Changing Appearance," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 1301–1307, 2020, doi: 10.1109/ICRA40945.2020.9196518.

[22] R. Dubé *et al.*, "SegMap: Segment-based mapping and localization using data-driven descriptors," *Int. J. Rob. Res.*, vol. 39, no. 2–3, pp. 339–355, 2020, doi: 10.1177/0278364919863090.