A comparative analysis of visual and point cloud-based place recognition methods in indoor environment.

Ivan Efremov
Robotics and Computer Vision
Innopolis University
Innopolis, Russia
i.efremov@innopolis.university

Ramil Khafizov
Robotics and Computer Vision
Innopolis University
Innopolis, Russia
r.khafizov@innopolis.university

Ramil Khusainov
Robotics and Computer Vision
Innopolis University
Innopolis, Russia
r.khusainov@innopolis.ru

***Abstract*—Place recognition plays a critical role in Simultaneous Localization and Mapping (SLAM) systems, as it helps in identifying previously visited locations and reducing localization errors. In this paper, we present a comprehensive evaluation of various place recognition methods, including DBoW2, SuperPoint + SuperGlue, Scan Context, and LoGG3D-Net, using data collected from indoor office and laboratory environments. Our evaluation focuses on the performance of these methods in terms of precision, recall, and F1 score, as well as their computational efficiency. We also explore the benefits of combining different methods to improve loop closure detection by leveraging complementary information from both visual and point cloud data. Our results indicate that vision-based methods generally exhibit better precision, while fusing methods leads to more robust performance. The evaluation code used in this study is available at our GitHub repository[1]. This comprehensive evaluation can serve as a valuable resource for researchers and practitioners working on SLAM and place recognition tasks, as well as inspire the development of novel approaches for improved performance and computational efficiency.**

***Index Terms*—SLAM, place recognition, loop closure, navigation, mobile robots**

I. Introduction

Simultaneous Localization and Mapping (SLAM) is a fundamental component in the development of autonomous vehicles and robotics. It enables robots to create a map of their environment while simultaneously localizing themselves within it. However, despite advancements in SLAM algorithms, there remains a persistent problem of error accumulation over time (Fig. **??**). This error can adversely affect the accuracy and robustness of the system, necessitating corrective measures.

One such corrective technique is loop closure, which is a specific instance of place recognition. Place recognition algorithms attempt to determine whether a robot has previously visited a specific location by comparing the current observation with historical data. Loop closure occurs when a match is found and the robot has completed a loop in its environment. In this case, the SLAM algorithm can correct the accumulated error by adjusting the robot's trajectory and map. Place recognition is an active area of research, and finding a suitable implementation for a specific use case can be challenging.

Recent surveys, such as those [1], [2] and [3], provide comprehensive overviews of common place recognition approaches. These approaches can generally be categorized into two main types: visual-based and point cloud-based place recognition.

Visual-based place recognition relies on image data, typically obtained from cameras, to identify previously visited locations. It leverages feature extraction and matching techniques to compare the current scene with those stored in memory. On the other hand, point cloud-based place recognition utilizes 3D data acquired from sensors such as LIDAR.

II. Goal

In this study, we focus on a specific hardware setup that includes a Livox LIDAR MID-70 sensor with a limited field of view and an Intel Depth Camera D435i. Despite the depth-sensing capabilities of the D435i, we will solely utilize its RGB information for the visual-based place recognition methods. The LIDAR and camera sensors are mounted on a handheld device, which makes our investigation particularly relevant for portable mapping and localization applications.

Our primary goal is to conduct a thorough comparison of various place recognition methods using data gathered from the small field of view LIDAR and RGB camera in indoor environments. Such environments often present unique challenges, such as fluctuating lighting conditions, clutter, and occlusions, which can significantly impact the accuracy and robustness of place recognition techniques.

We will perform all tests and evaluations exclusively in indoor settings, taking into account factors like computational efficiency, robustness to dynamic changes, and resilience to perceptual aliasing. By exploring these factors, we aim to provide valuable insights and recommendations for the deployment of effective place recognition methods in indoor SLAM applications that employ RGB cameras and/or small field of view LIDAR.

[1]https://github.com/4ku/Place-recognition-evaluation

III. Problems

In this section, we outline the primary and specific challenges associated with place recognition in indoor environments using the given LIDAR and camera setup.

The primary challenges for place recognition consist of five key aspects [2]: 1) Perceptual aliasing, 2) Appearance change, 3) Viewpoint difference, 4) Generalization ability, and 5) Efficiency and robustness.

- **Perceptual aliasing**: Indoor environments often contain similar appearances for distinct places, resulting in perceptual aliasing (Fig. 1). This issue can stem from repetitive patterns, uniform textures, or symmetric structures commonly found in man-made environments. Addressing perceptual aliasing is critical for improving place recognition performance in both LIDAR and camera-based systems.
- **Appearance change**: Place recognition in indoor environments can be particularly challenging due to appearance changes caused by conditional factors like varying lighting conditions and structural factors such as clutter, occlusions, and repetitive structures. Conditional changes impact visual observations over time, while structural changes affect both vision and LiDAR sensors.
- **Viewpoint difference**: Place recognition methods must be able to account for varying perspectives, as observations in the same place may have different patterns when viewed from different angles. This is especially important in long-term navigation.
- **Generalization ability**: An effective place recognition method should be capable of recognizing unseen environments. Since real-world scenarios are infinite and the same place can be presented under different environmental conditions, it is impossible to collect all possible combinations of place datasets at once.
- **Efficiency and robustness**: For real-time applications, the performance of the place recognition method should be computationally efficient and robust, capable of operating on a embedded devices without causing significant delays.

Problems with point cloud place recognition:

- Livox LIDAR has a small field of view, which may limit its effectiveness in certain place recognition approaches, as most state-of-the-art methods are designed for 360-degree LIDAR sensors.
- Most state-of-the-art place recognition approaches are tailored for outdoor environments and autonomous cars. Applying these approaches to indoor settings may lead to unexpected challenges and require significant adaptations.

Problems with visual place recognition:

- Visual place recognition methods are sensitive to variations in lighting conditions, which are common in indoor environments. Shadows, reflections, and artificial lighting can introduce inconsistencies between image data, making it challenging to identify previously visited locations.
- Indoor environments may contain dynamic elements such as moving objects, people, or temporary occlusions. These factors can negatively impact the performance of visual place recognition algorithms and lead to false positive or false negative loop closures.
- Visual place recognition methods may face limitations in handling changes in scale and perspective when relying solely on RGB information. This could lead to reduced accuracy and robustness.

IV. Overview of Algorithms

A. Point cloud place recognition methods

Recent years have witnessed a surge in point cloud place recognition methods, many of which are documented in a

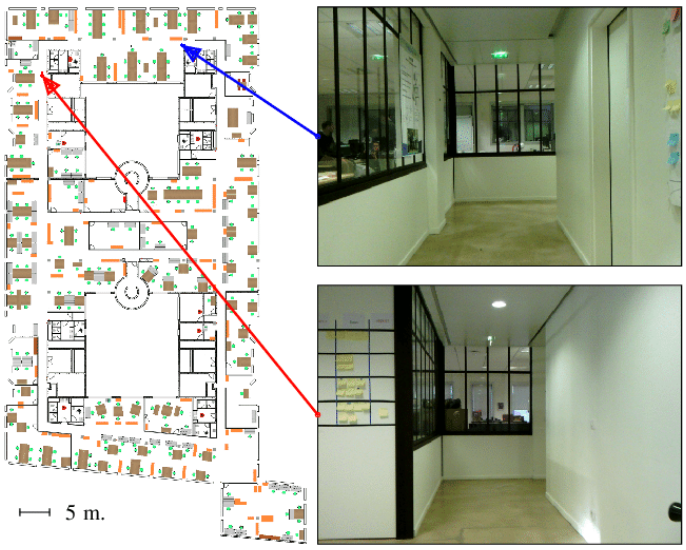

Fig. 1. Example of perceptual aliasing [4].

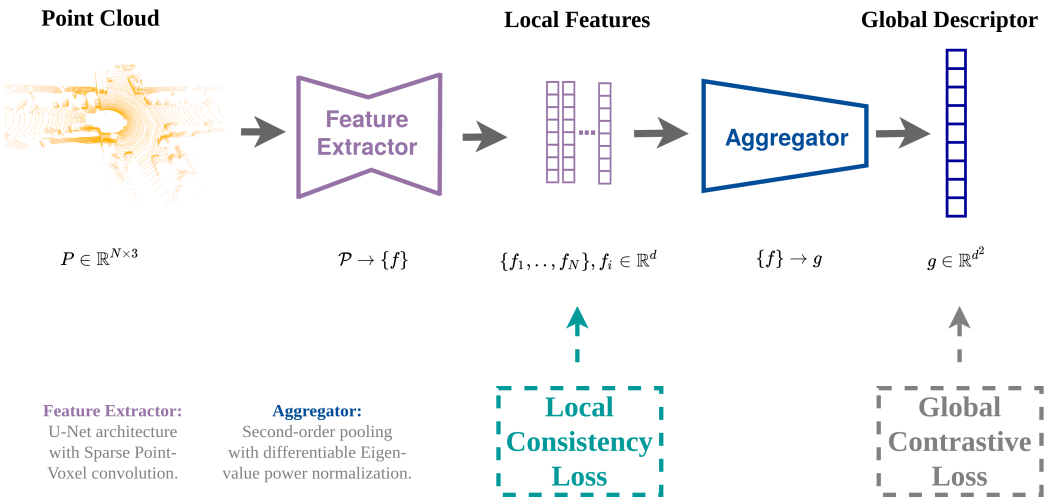

Fig. 2. LoGG3D-Net [5] pipeline for global descriptor retrieval.

comprehensive GitHub list[2]. This section briefly overviews common approaches, their pros and cons, and their suitability for small field of view LIDARs like Livox LIDAR.

State-of-the-art OverlapTransformer [6] excels in performance but is limited to 360-degree LIDAR sensors. It employs a lightweight neural network for fast execution and yaw-angle-invariant architecture, enhancing place recognition.

"Robust Place Recognition using an Imaging Lidar" [7] maps point clouds to images but is not well-suited for Livox LIDAR sensors. This rotation-invariant method generates an intensity image from intensity readings, encoding it into a bag-of-words vector for place recognition.

Scan Context [8], an older yet still relevant method, records visible space directly and calculates similarity scores without relying on histograms or prior training. It uses a two-phase search algorithm for efficient loop detection. Although not specifically designed for small field of view LIDARs, it can be suitable for such sensors as well.

LoGG3D-Net [5] is using local consistency loss to improve 3D place recognition performance. It achieves state-of-the-art results on KITTI and MulRan [9] datasets, operating in near real-time. Like Scan Context, it is not specifically designed for small field of view LIDARs but can be used with such sensors also.

In this paper, we focus on Scan Context and LoGG3D-Net due to their relevance, performance, and potential compatibility with small field of view Livox LIDAR sensors. We will investigate their suitability for such sensors, compare these methods with visual place recognition methods, and evaluate their performance using our dataset.

B. Visual place recognition methods

Visual place recognition (VPR) methods offer cost-effective alternatives to LIDAR sensors for identifying locations within an environment. VPR methods' affordability and strong performance have made them popular for place recognition tasks.

[2]https://github.com/kxhit/awesome-point-cloud-place-recognition

Bag of Visual Words (BoVW) [10], a well-established VPR method, quantizes local image features into discrete visual words by clustering them into a visual vocabulary. An image is represented as a histogram of these visual words, forming a global image descriptor. BoVW's performance can be influenced by local features, clustering algorithms, and visual vocabulary size.

NetVLAD [11] integrates the VLAD technique into a CNN architecture, extracting local features and aggregating them into a compact global descriptor. This method excels in large-scale place recognition tasks, even under challenging conditions.

SuperPoint [12] simultaneously detects keypoints and extracts local descriptors from input images, using a self-supervised training approach for geometrically stable feature representation. SuperGlue [13], a graph neural network-based approach for feature matching, is used in conjunction with SuperPoint for robust and efficient keypoint correspondence.

In this paper, we focus on BoVW and the combination of SuperPoint and SuperGlue. Chosen for their performance, relevance, these methods will be compared and evaluated using our dataset.

V. Datasets

We collected data using a handheld device equipped with a Livox LIDAR MID-70 and an Intel Depth Camera D435i. The SLAM algorithm was then employed to estimate the trajectories, or odometry. While the primary objective of place recognition is to reduce errors, our dataset exhibits a relatively small error due to the short trajectories involved. This characteristic makes the dataset suitable for evaluating various place recognition methods.

Our dataset is stored in a rosbag file containing three topics: one topic with RGB images from the camera, another topic with point cloud data from the Livox LIDAR, and the last topic containing the odometry information estimated by the SLAM algorithm.

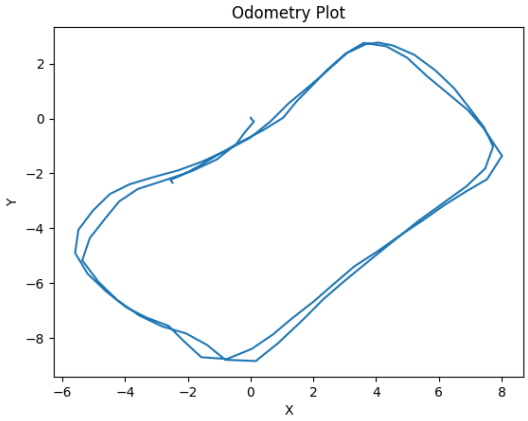

Fig. 3. Trajectory of data collected around an office room.

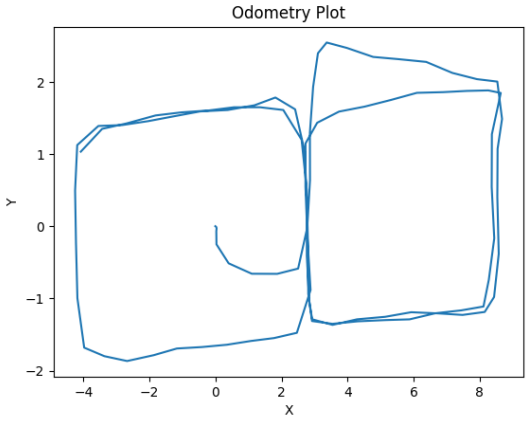

Fig. 4. Trajectory of data from a laboratory environment.

The data was collected at Innopolis University, with two distinct rosbag files representing different environments. The first rosbag file contains data collected in and around an office room, featuring two loops (Fig. 3). The second rosbag file captures data from a laboratory environment, with a figure-eight-shaped trajectory that includes two loops as well (Fig. 4). The office and lab environments possess distinct geometries and visual appearances, making them suitable for use as training and testing data for place recognition methods.

VI. Model Preparation

Initially, we fine-tune each place recognition method on the train rosbag collected near the office. The fine-tuning process involves manually determining the threshold for each model that yields the best place recognition results for the train dataset. Our primary focus is on achieving the highest possible precision on the train dataset, as precision plays a crucial role in the effectiveness of place recognition methods. The importance of precision is further discussed in the end of the section VII.

For the Bag of Words (BoW) method, we use the DBoW2 implementation[3]. Before evaluation, we create a BoW database for improved predictions using another rosbag collected near the office. We then use the train dataset to find the appropriate score threshold that provides the best results.

For the SuperPoint + SuperGlue method, we utilize pre-trained indoor weights based on ScanNet data, available at source repository[4]. The model computes the number of matched features between two images. Following the fine-tuning process, we identify the optimal threshold for the number of matched features, signifying that the two images correspond to the same location.

For point cloud-based methods, we use the Livox MID-70 LIDAR for testing. The point clouds generated by the MID-70 are sparse petals, making single clouds extremely challenging to recognize, even for human observers. As none of the methods succeeded with individual clouds, we accumulated clouds into either the global frame or the local frame of the first cloud in the batch. Each accumulated cloud contains 10 single clouds, which corresponds to the number of point clouds collected per second by the Livox MID-70 LIDAR. We chose 10 clouds because this quantity allows for more easily recognizable structures to the human eye. Additionally, all point clouds were transformed into local coordinates by translating and rotating them based on the first point cloud's translation and rotation to improve algorithm performance.

The Scan Context method[5] does not require pretrained weights. The algorithm computes the distance between Scan Context descriptors. If the distance is small between two descriptors, the corresponding points are considered to be the same.

The LoGG3D-Net method[6] is pretrained on the outdoor MulRan dataset but still performs quite well in indoor environments. However, it's better to train on indoor data to improve results further. Similar to Scan Context, LoGG3D-Net calculates the distance between descriptors.

VII. Models Evaluation

We define two points on a trajectory to be the same if they are within a 3-meter radius of each other. Additionally, due to the Livox LIDAR MID-70's limited 70-degree field of view, we account for device orientation by considering the device to be looking in the same direction if the angular distance is less than 45 degrees. Consequently, we examine two cases:

1) When two points on the trajectory are close to each other (within 3 meters)
2) When two points on the trajectory are close to each other and oriented in the same direction (within 45 degrees)

During the evaluation process, we collected data frames that include odometry, image, and point cloud data from testing data in the laboratory environment. Each data frame

[3]https://github.com/kxhit/awesome-point-cloud-place-recognition
[4]https://github.com/magicleap/SuperGluePretrainedNetwork
[5]Source code: https://github.com/irapkaist/scancontext
[6]Source code: https://github.com/csiro-robotics/LoGG3D-Net

was compared with all previously collected data frames to make predictions for loop closure candidates using models fine-tuned on the train data. Subsequently, we calculated performance metrics such as precision, recall, and F1 score. The results are presented in Tables I and II.

TABLE I
MODELS EVALUATION ON TEST DATA (LABORATORY ENVIRONMENT). TRUE POSITIVE HERE IS WHEN TWO POINTS ARE CLOSE TO EACH OTHER (WITHIN 3 METERS).

Method	Precision	Recall	F1 Score
DBoW2	0.946	0.137	0.240
SuperPoint + SuperGlue	0.971	0.307	0.466
Scan Context	0.948	0.196	0.325
LoGG3D-Net	0.792	0.122	0.211

TABLE II
MODELS EVALUATION ON TEST DATA (LABORATORY ENVIRONMENT). TRUE POSITIVE HERE IS WHEN TWO POINTS ARE CLOSE TO EACH OTHER (WITHIN 3 METERS) AND ORIENTED IN THE SAME DIRECTION (WITHIN 45 DEGREES).

Method	Precision	Recall	F1 Score
DBoW2	0.941	0.324	0.482
SuperPoint + SuperGlue	0.971	0.728	0.832
Scan Context	0.708	0.348	0.467
LoGG3D-Net	0.782	0.285	0.418

Given the Livox LIDAR MID-70's limited field of view, results that take angle consideration into account are more reliable. As such, it is recommended to rely on these results, as they provide a more accurate representation of the model's performance in recognizing places.

A. Combining methods

In addition to evaluating the performance of each method independently, we also explore the potential benefits of combining different methods to improve loop closure detection. For this analysis, we consider two points to be a loop candidate only if both methods under consideration identify them as a loop candidate simultaneously. The results of this combination approach can be seen in Tables III and IV below.

TABLE III
COMBINED MODELS EVALUATION ON TEST DATA. TRUE POSITIVE HERE IS WHEN TWO POINTS ARE CLOSE TO EACH OTHER (WITHIN 3 METERS). **NOTE**: S.P. = SUPERPOINT, S.G. = SUPERGLUE.

Method	Prec.	Recall	F1
DBoW2 + LoGG3D	0.977	0.060	0.113
DBoW2 + Scan C.	0.990	0.071	0.133
(S.P. + S.G.) + Scan C.	1.0	0.132	0.234
(S.P. + S.G.) + LoGG3D	0.970	0.114	0.204

The primary metric of interest is precision, as it is crucial to ensure that the place recognition model does not produce incorrect loop candidates. Erroneous loop candidates can exacerbate SLAM errors rather than mitigating them.

By fusing these approaches, we achieve a more robust loop closure detection performance in terms of precision. The combined models exhibit higher precision values, which demonstrates the effectiveness of leveraging complementary information from both visual and point cloud data. The fusion of methods compensates for the individual limitations of each technique, resulting in enhanced overall performance.

TABLE IV
COMBINED MODELS EVALUATION ON TEST DATA. TRUE POSITIVE HERE IS WHEN TWO POINTS ARE CLOSE TO EACH OTHER (WITHIN 3 METERS) AND ORIENTED IN THE SAME DIRECTION (WITHIN 45 DEGREES). **NOTE**: S.P. = SUPERPOINT, S.G. = SUPERGLUE.

Method	Prec.	Recall	F1
DBoW2 + LoGG3D	0.977	0.142	0.248
DBoW2 + Scan C.	0.990	0.169	0.289
(S.P. + S.G.) + Scan C.	1.0	0.314	0.478
(S.P. + S.G.) + LoGG3D	0.970	0.270	0.423

VIII. SPEED PERFORMANCE

We evaluated the computational efficiency of each method by analyzing their execution time on a testing data consisting of 100 frames (i.e., 100 seconds of data). These performance evaluations were conducted on an AMD Ryzen 7 5800H CPU and NVIDIA GeForce RTX 3050 Ti Laptop GPU. The results, presented in seconds for consistency, can be found at Table V.

TABLE V
EXECUTION TIME PERFORMANCE OF EACH METHOD ON 100 FRAMES (I.E., 100 SECONDS OF DATA).

Method	Processing unit(s)	Total Duration (s)
DBoW2	CPU	2.65
SuperPoint + SuperGlue	CPU + GPU	357.96
Scan Context	CPU	0.34
LoGG3D-Net	CPU + GPU	10.39

IX. CONCLUSION

In this study, we evaluated various place recognition methods, including DBoW2, SuperPoint + SuperGlue, Scan Context, and LoGG3D-Net, in an indoor laboratory environment using both visual and point cloud data. The results indicate that vision-based methods, DBoW2 and SuperPoint + SuperGlue, exhibit better precision in detecting loop closures compared to point cloud-based methods. This outcome can be attributed to several factors:

- The point cloud methods were initially designed for 360-degree LIDAR systems and may not be well-suited for the limited field of view provided by the Livox LIDAR MID-70.
- Scan Context, being a relatively simple and older method, may not offer the same level of performance as more recent, sophisticated algorithms.
- LoGG3D-Net, while pretrained on outdoor data, still performs reasonably well in indoor environments. However, additional training on indoor data could potentially improve its performance.

Additionally, we explored the benefits of combining different methods to enhance loop closure detection performance.

The fusion of these approaches compensates for the individual limitations of each technique, yielding a more robust performance in terms of precision. This demonstrates the effectiveness of leveraging complementary information from both visual and point cloud data.

Furthermore, we evaluated the computational efficiency of each method, which is a crucial factor in real-time SLAM applications. Among the four methods, Scan Context demonstrated the fastest execution time, followed by DBoW2, LoGG3D-Net, and SuperPoint + SuperGlue.

In conclusion, this study highlights the importance of selecting appropriate place recognition methods for loop closure detection in SLAM systems. By carefully considering the strengths and weaknesses of each method and potentially combining complementary approaches, researchers and practitioners can design more robust and efficient SLAM systems that are better suited to the specific requirements of various applications.

REFERENCES

[1] T. Barros, R. Pereira, L. Garrote, C. Premebida, and U. J. Nunes, "Place recognition survey: An update on deep learning approaches," *arXiv preprint arXiv:2106.10458*, 2021.

[2] P. Yin, S. Zhao, I. Cisneros, *et al.*, *General place recognition survey: Towards the real-world autonomy age*, 2022. arXiv: 2209.04497 [cs.RO].

[3] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *arXiv preprint arXiv:2103.06443*, 2021.

[4] M. Nowakowski, C. Joly, S. Dalibard, N. Garcia, and F. Moutarde, "Topological localization using wifi and vision merged into fabmap framework," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 2017, pp. 3339–3344.

[5] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes, "Logg3d-net: Locally guided global descriptor learning for 3d place recognition," in *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 2215–2221.

[6] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition," *IEEE Robotics and Automation Letters*, 2022.

[7] T. Shan, B. Englot, F. Duarte, C. Ratti, and D. Rus, "Robust place recognition using an imaging lidar," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 5469–5475.

[8] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Oct. 2018.

[9] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "Mulran: Multimodal range dataset for urban place recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 6246–6253.

[10] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012, ISSN: 1552-3098. DOI: 10.1109/TRO.2012.2197158.

[11] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, *Netvlad: Cnn architecture for weakly supervised place recognition*, 2016. arXiv: 1511.07247 [cs.CV].

[12] D. DeTone, T. Malisiewicz, and A. Rabinovich, *Superpoint: Self-supervised interest point detection and description*, 2018. arXiv: 1712.07629 [cs.CV].

[13] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, *Superglue: Learning feature matching with graph neural networks*, 2020. arXiv: 1911 . 11763 [cs.CV].