Super-BoW Algorithm for Place Recognition in Indoor Environments

Kesse Jonatas de Jesus*, Lucas Marchesan Silva*, Gustavo Arthur Dutra*,
Daniel Fernando Tello Gamarra*, Anselmo Rafael Cukla*, Thássio Gomes Silva*, Felipe Oliveira Silva†,
Rodrigo Silva Guerra‡, Paulo Lilles Jorge Drews-Jr‡, Lucas Benedetti Viana Cordova‡
*Federal University of Santa Maria, Santa Maria, Brazil
†Federal University of Lavras, Lavras, Brazil
‡Federal University of Rio Grande, Rio Grande, Brazil

***Abstract*—Visual place recognition in indoor environments remains a challenging problem due to illumination changes, perceptual aliasing, and the need for efficient large-scale image retrieval. This paper presents Super-BoW, a Bag-of-Words (BoW) retrieval pipeline designed for image matching using learned local descriptors. The proposed method combines SuperPoint for keypoint detection and 256-dimensional descriptor extraction with a MiniBatchKMeans visual vocabulary to generate compact and discriminative BoW histograms. The system operates in two stages: a training stage, where the visual vocabulary is learned from indoor images, and a testing stage, where query images are encoded and compared using an efficient BallTree search. Image preprocessing techniques such as histogram equalization and Contrast Limited Adaptive Histogram Equalization (CLAHE) are applied to improve descriptor robustness under illumination variations. Real-world experiments demonstrate that Super-BoW achieves 100% Top-5 retrieval accuracy while maintaining low computational cost.**

***Index Terms*—Image retrieval, Bag-of-Words, SuperPoint, visual vocabulary, MiniBatchKMeans**

I. Introduction

Imagine a robot traversing a long, monotonous hallway on a university campus or a drone exploring the interior of a building: its ability to navigate safely depends on knowing where it is.

However, real-world navigation imposes severe challenges, such as extreme variations in viewpoint, illumination changes, and the presence of dynamic objects, which can drastically alter the appearance of a scene over time [18]. To maintain autonomy in such conditions, robust feature extraction and efficient retrieval techniques are required to ensure reliable performance in long-term operations [19].

Deep neural networks have increasingly filled this role in visual place recognition and SLAM systems [1], [6], [17]. The SuperPoint model, for instance, is a fully convolutional architecture that computes keypoints and high-level descriptors for multiview geometry problems in a single pass [14]. Trained with homographic adaptations on image collections such as MS-COCO, SuperPoint detects a much richer set of repeatable points than classical detectors and achieves state-of-the-art results in homography estimation. However, for these descriptors to be used in large image databases, they must be aggregated into compact representations. This is where *Bag-of-Words* (BoW) models come in [4], [7], [12]: an image is treated as a "document," local features are detected, described as numerical vectors, and then algorithms such as k-means [13] are used to cluster these vectors and form a visual vocabulary. Each image is then represented by a histogram of frequencies of these "visual words." To accelerate nearest-neighbor queries, indexing structures like Ball Trees store the data in a tree format and reduce search time [15].

In this context, we propose *SuperBoW*: an architecture that combines SuperPoint descriptors with a visual vocabulary generated by MiniBatchKMeans [16] and represents each image as a BoW histogram. Retrieval is accelerated by a Ball Tree, enabling near real-time searches in large databases. We evaluate the method on an indoor image set collected at the Federal University Technology Center, which presents variations in texture, contrast, and lighting. The results show that SuperBoW achieves 100 % Top-5 accuracy and computational performance compatible with mobile robotics applications.

II. Related Works

Place recognition has been extensively studied in the computer-vision and robotics communities as a fundamental problem for visual localization, navigation, and loop-closure detection. A comprehensive overview of the field is presented in [1], which analyzes both traditional place-recognition pipelines and modern Convolutional Neural Networks (CNN)-based approaches, discussing the challenges associated with appearance changes, perceptual aliasing, and viewpoint variations.

In [17], the authors investigate visual localization performance from a place-recognition perspective, highlighting how feature selection, descriptor robustness, and matching strategies directly influence recognition accuracy in both indoor and outdoor environments. A novel image-sequence-based framework for mobile robot place recognition was also proposed, which models a place using an image sequence and leverages the Echo State Network (ESN) for robust place modeling and segmentation [2]. Furthermore, a cognition-inspired framework coined Memorable Maps was introduced, redefining the concept of a place in visual place recognition

979-8-3195-1715-9/26/$31.00 ©2026 IEEE

by filtering input frames based on memorability, staticity, and entropy to discard confusing or non-salient images [3].

BoW representations have also been explored for place recognition beyond purely visual descriptors. In [4], a BoW-based strategy is employed to enable robust place matching under viewpoint invariance, demonstrating that compact vocabularies remain effective for large-scale retrieval tasks. More recent works investigate how multi-modal cues can support reliable place discrimination in environments subject to perceptual aliasing and structural repetition [5].

Several approaches leverage learned representations to improve robustness in place recognition under challenging conditions. In [6], deep feature extraction combined with structured matching strategies is shown to improve correspondence quality, which directly benefits image-based place recognition. The Trajectory Segmentation-based Online Bag of Words(TSO-BoW) method [7] further demonstrates that BoW models can sustain accurate long-term place recognition with constant query time and bounded memory requirements. Additional works explore robustness to illumination changes, motion artifacts, and image degradation, emphasizing their impact on reliable place recognition performance [8].

Indoor environments pose particular challenges for place recognition due to repetitive structures, limited geometric variation, and illumination changes. Lee et al. [9] analyze indoor visual localization scenarios and demonstrate how sensing modalities influence recognition reliability. Similarly, Cramariuc et al. [10] discuss modular visual mapping frameworks that support systematic evaluation of place-recognition components across different environments. A comprehensive experimental study [11] further examines how acquisition conditions affect indoor visual localization accuracy, reinforcing the importance of robust place-recognition strategies.

Our approach targets the place-recognition stage specifically by combining learned SuperPoint descriptors with a lightweight BoW representation optimized via MiniBatchK-Means. While conceptually aligned with BoW-based loop-closure techniques and learned-feature localization methods, the proposed SuperBoW system differs by emphasizing efficient large-scale image retrieval using a compact histogram representation and BallTree acceleration. This fills a gap in the literature by demonstrating that learned keypoints can be effectively integrated into classical BoW pipelines to achieve real-time indoor place-recognition performance.

III. Theoretical Background

This section presents the theoretical foundation necessary for the development and understanding of the proposed work. Initially, the BoW model and the K-Means clustering algorithm are discussed, which are essential techniques for the representation and quantization of visual features. Next, the SuperPoint architecture is addressed, a deep learning-based method for the detection and description of interest points. Finally, the BallTree data structure is presented, used to optimize search and indexing in multidimensional spaces.

A. Bag-of-Words

The BoW model represents images through histograms of "visual words," formed after detecting interest points, extracting descriptors, and grouping them into a visual vocabulary. In [12], the authors explain that this approach allows for transforming local features into a simple and effective representation for image annotation and classification tasks.

B. K-Means

The K-means algorithm is an iterative partitioning procedure that divides N objects into K disjoint clusters, with the objective of maximizing the expected similarity between data items and the centroids of their associated clusters. Most variants of this algorithm use Euclidean distance as a metric, applying the minimum distance rule to assign each data point to its nearest centroid [13]. The formulation of the squared Euclidean distance between a data point x and a centroid c, fundamental for the calculation of the intra-cluster squared error, is given by:

$$d(x, c) = ||x - c||^2 = \sum_{j=1}^{d} (x_j - c_j)^2 \tag{1}$$

where d represents the dimensionality of the data.

A commonly used scalable variant of this method is *Mini-BatchKMeans* [16]. Instead of processing the entire dataset at each iteration, MiniBatchKMeans updates the centroids using small random subsets of the data, known as mini-batches. These incremental updates significantly reduce computational cost and memory usage while still guiding the centroids toward good cluster positions. Although the resulting clusters may be slightly less precise than those produced by standard K-Means, this approach is considerably faster and well suited for large-scale or high-dimensional datasets, such as image descriptors in visual Bag-of-Words systems.

C. SuperPoint

SuperPoint is a fully convolutional neural network architecture proposed for the self-supervised detection and description of interest points [14]. Unlike patch-based approaches, the model operates on full-sized images and jointly calculates the location of interest points at the pixel level and their associated descriptors in a single forward pass [14]. To enable training without manual human annotation, the authors introduced "Homographic Adaptation," a multi-scale and multi-homography technique that increases point detection repeatability and performs domain adaptation, allowing the model initially trained on synthetic shapes to generalize to real images [14]. Each overlaid point in Fig. 1 represents a SuperPoint keypoint detected by the network, indicating locations of high visual saliency learned in a self-supervised manner. These keypoints are associated with learned descriptors, enabling robust matching across different views of the same environment.

Fig. 1. Features extracted in an image with the SuperPoint algorithm.

D. BallTree

Balltrees are geometric data structures organized as complete binary trees, where each node is associated with a ball (hypersphere) in an n-dimensional Euclidean space [15]. Unlike *k-d trees*, sibling regions in a Balltree can intersect and do not need to partition the entire space; the ball of an internal node is defined as the smallest sphere that contains the balls of its children [15]. Omohundro [15] compares five construction algorithms, concluding that the *bottom-up* approach generally produces trees of better quality (minimizing the total volume of the balls), although it presents the highest construction time, while the *k-d tree* is faster but adapts more poorly to data with a clustered structure.

Fig. 2 illustrates the conceptual organization of a BallTree. The left side of the figure shows the geometric interpretation in feature space, where nested and possibly overlapping hyperspheres enclose subsets of data points. Each hypersphere represents a node in the tree and bounds a group of feature vectors. The right side depicts the corresponding tree structure, highlighting the hierarchical subdivision of the data into progressively smaller balls until leaf nodes are reached.

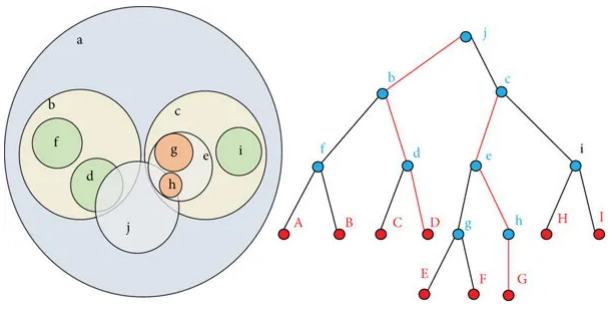

Fig. 2. Balltree structure.

IV. Materials and Methods

The experimental framework used in this work, as well as the proposed image–retrieval architecture based on SuperPoint descriptors and a BoW approach, is described below.

A. Dataset

The dataset employed in this study consists of indoor images collected at the University Technology Center, covering three differents areas. All images are provided in high resolution and exhibited substantial variability in illumination, contrast, and local texture density — characteristics that make them suitable for evaluating place-recognition methods under challenging real-world conditions. Figures 3 to 5 show dataset sequences named corridor1, corridor2, and sala1103, respectively. The dataset consists of 5,468 images acquired with a resolution of 1920×1080 pixels. Following an 80/20 split, 4,374 images were used for vocabulary training and histogram generation, while 1,094 images served as query images for evaluation.

A training subset of the Federal University Dataset was used to build the visual vocabulary, while a separate subset was reserved for evaluation. Each image was processed using the SuperPoint network to extract local keypoints and 256-dimensional descriptors. All descriptors from the training images were concatenated and clustered using MiniBatchK-Means to generate the visual vocabulary used in the Bag-of-Words representation.

For the evaluation stage, each query image was converted into a BoW histogram using the trained vocabulary. Ground-truth correspondences between query images and their correct matches were defined through a manually verified reference list. This list was used exclusively to compute the Top-K retrieval accuracy. Top-K retrieval accuracy measures the percentage of query images for which the correct reference image is retrieved within the top K ranked results.

Fig. 3. Dataset images of the corredor1 sequence.

Fig. 4. Dataset images of the corredor2 sequence.

Fig. 5. Dataset images of the sala1103 sequence.

B. Super-BoW Architecture

Figure 6 presents the complete SuperBoW pipeline divided into two stages: a training stage, where the visual vocabulary is learned, and a testing stage, where a query image is processed and matched against the database.

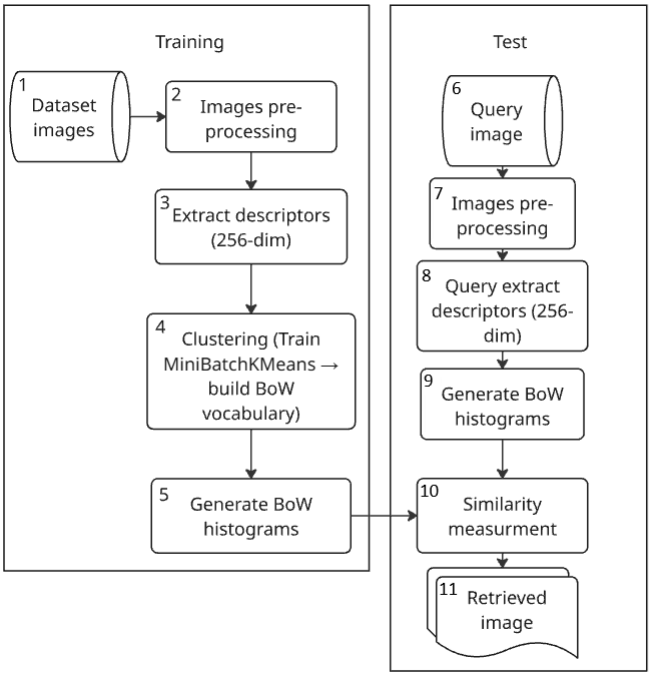

Fig. 6. Overview of the proposed SuperBoW architecture.

In the training stage, the process begins with (1) Dataset Images, which serve as input to the pipeline. Each image is enhanced through (2) Image Pre-processing, using histogram equalization and CLAHE to improve global and local contrast. Next, SuperPoint is applied to (3) Extract Descriptors (256-dim), generating keypoints and their 256-dimensional learned descriptors.

All descriptors from the training set are aggregated and clustered during (4) Clustering MiniBatchKMeans, producing the visual vocabulary. Using this vocabulary, each training image is converted into a normalized BoW representation in (5) Generate BoW Histograms. which is used at retrieval time.

In the testing stage, the system receives a (6) Query Image, which undergoes the same (7) Image Pre-processing to ensure consistency. SuperPoint is again used to (8) Query Extract Descriptors (256-dim), and the resulting descriptors are mapped into the vocabulary to (9) Generate BoW Histogram for the query. A nearest-neighbor search then performs (10) Similarity Measurement using a BallTree to efficiently compare the query histogram with the database. The connection between Blocks (5) and (10) highlights that similarity measurement is performed by comparing the BoW histogram of the query image with the BoW histograms generated from the training images. Finally, the system returns the (11) Retrieved Image, corresponding to the closest match or the Top-K candidates used for evaluation.

Algorithm 1 BoW Histogram Generation

Input: Set of image descriptors $\mathcal{D} = \{D_1, \ldots, D_N\}$, visual vocabulary $\mathcal{V}$ with K clusters
Output: Set of normalized BoW histograms $\mathcal{H}$
for each image descriptor set $D_i \in \mathcal{D}$ **do**
  if $|D_i| > 0$ **then**
    Assign each descriptor in D_i to the nearest visual word in $\mathcal{V}$
    Build histogram h_i by counting visual word occurrences
    Normalize h_i
  else
    Set h_i as a zero vector of dimension K
  end if
  Add h_i to $\mathcal{H}$
end for
Return $\mathcal{H}$

Algorithm 1 describes the generation of BoW histograms. For each image, local descriptors are assigned to the nearest visual words in the learned vocabulary, and the resulting histogram is normalized to ensure scale invariance across images as shown the blocks (5) and (9) in Figure 6.

The system was implemented in Python, using PyTorch for SuperPoint inference, OpenCV for image preprocessing, and scikit-learn for MiniBatchKMeans clustering, Bag-of-Words generation, and nearest-neighbor retrieval with BallTree.

C. Preprocessing and Descriptor Extraction

Each image was resized to 640×480, converted to grayscale, and enhanced using CLAHE and global histogram equalization. The SuperPoint network extracted up to 1,024 keypoints per frame, each represented by a 256-dimensional descriptor. Across all training images, a total of 3,178,958 descriptors were collected and concatenated to train the MiniBatchKMeans model responsible for generating the BoW vocabulary.

The preprocessing and vocabulary-building stage required approximately 6 minutes on the evaluation hardware (Intel i3-10100F CPU and NVIDIA GTX 1660 SUPER GPU). The resulting artifacts included the visual vocabulary (`vocab.npy`), BoW histograms for all images (`histograms.npy`), and the trained clustering model (`MiniBatchKMeans_model.pkl`).

D. Evaluation Protocol

During testing, each query image was processed by:

1) Extracting SuperPoint descriptors,
2) Converting descriptors into a BoW histogram,
3) Querying the BallTree structure to retrieve the Top-N nearest neighbors.

A match was considered correct if any of the retrieved Top-N images corresponded to the true ground-truth location of that query. Because the dataset contains highly repetitive corridors, Top-1 identification is challenging, whereas Top-5 provides a more robust measure of retrieval performance.

V. Results

This section presents the experimental results obtained using the proposed SuperBoW architecture. All experiments were conducted using the Federal University indoor dataset, which contains three visually distinct locations: `corredor_1_andar`, `corredor_2_andar`, and `sala_1103`.

A. Qualitative Analysis

Qualitative inspection of retrieved matches revealed that:

- The system performed strongly in textured indoor scenes, particularly when the frames included multiple structural edges, doors, signs, or objects.
- Retrieval errors in the Top-1 evaluation occurred primarily in long uniform corridors, where frames exhibit near-identical surface patterns and low geometric variance.
- The CLAHE preprocessing step noticeably improved descriptor consistency under low illumination, resulting in more stable clustering behavior during vocabulary training.

Figure 7 and 8 show an example of the visual interface during the simulation of the sequence corredor2 and sala1103 respectively, highlighting retrieved matches and histogram comparisons.

Fig. 7. Example of the system visualization interface during the SuperBoW simulation of the sequence corredor2.

B. Computational Performance

The full evaluation of all 1,094 test frames required 2 minutes and 40 seconds, corresponding to an average processing time of 0.147 seconds per query which is approximately 6.8 queries per second. This includes descriptor extraction, histogram generation, and nearest-neighbor search. The BallTree

Fig. 8. Example of the system visualization interface during the SuperBoW simulation of the sequence sala1103.

structure contributed significantly to fast retrieval, with per-query lookup times on the order of a few milliseconds. Most of the computational cost originated from SuperPoint feature extraction.

C. Quantitative Results

Table I summarizes the accuracy metrics obtained on the 1,094 test images.

TABLE I
Retrieval Accuracy on the Indoor Dataset

Metric	Value
Top-1 Accuracy	62.25%
Top-5 Accuracy	**100.00%**

The Top-1 accuracy of 62.25% reflects the visual ambiguity inherent to indoor corridor environments, where many frames contain similar textures, illumination, and geometric structures. Nevertheless, the system correctly identified the true location within the Top-5 retrieved candidates for **all** test images, demonstrating the robustness of the BoW representation in constrained environments.

D. Discussion

The results demonstrate that the SuperBoW architecture is capable of reliably retrieving the correct indoor location under challenging appearance variations. The 100% Top-5 accuracy indicates that the BoW representation captures sufficient discriminative information even in environments with strong structural repetition. The fast processing rate further confirms the suitability of the approach for real-time or large-scale localization applications.

Overall, the combination of contrast-enhancing preprocessing, SuperPoint descriptors, and a BoW histogram representation proved to be both effective and computationally efficient.

E. Retrieval Accuracy

Across all query images, the system demonstrated stable performance with fast inference and consistent retrieval quality. Histogram representations generated from SuperPoint descriptors were found to be discriminative even in regions with repetitive textures or low structural complexity.

For each query, the BallTree structure returned the Top-5 most similar images based on histogram distance. A match was considered correct if any of the Top-5 returned images

corresponded to one of the true references in the ground-truth list.

Qualitatively, the system produced correct matches even when the visual appearance differed due to illumination variations or local contrast changes. This demonstrates the effectiveness of combining SuperPoint descriptors with a BoW aggregation strategy.

VI. Conclusions

The proposed visual localization system based on SuperPoint and BoW demonstrated that learned local descriptors can be effectively used for large-scale image retrieval. The approach achieved a Top-5 accuracy of 100%, indicating that the correct match is consistently included among the top retrieved candidates. The Top-1 accuracy of 62.25% reflects the challenges of viewpoint variations and illumination changes present in the employed indoor dataset.

The results show that the combination of SuperPoint descriptors with a MiniBatchKMeans-based visual vocabulary provides a fast and scalable matching strategy, with an average processing time of 0.147 seconds per query image. These findings confirm that the method is suitable for real-time or near real-time visual localization scenarios.

Future work includes extending the SuperBoW approach to a full visual odometry pipeline and, potentially, to a complete visual SLAM system. The authors gratefully acknowledge the support of CNPq.

Acknowledgment

References

[1] Z. Zeng, J. Zhang, X. Wang, Y. Chen, and C. Zhu, "Place recognition: An overview of vision perspective," *Sensors*, vol. 20, no. 14, p. 4037, Jul. 2020, doi: 10.3390/s20144037.

[2] J. Yuan *et al.*, "A novel approach to image-sequence-based mobile robot place recognition," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 51, no. 9, pp. 5377–5391, Sept. 2021, doi: 10.1109/TSMC.2019.2956321.

[3] M. Zaffar, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "Memorable maps: A framework for re-defining places in visual place recognition," *IEEE Trans. Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7355–7369, Dec. 2021, doi: 10.1109/TITS.2020.3001228.

[4] Y. Cui, X. Chen, Y. Zhang, J. Dong, Q. Wu, and F. Zhu, "BoW3D: Bag of words for real-time loop closing in 3D LiDAR SLAM," *IEEE Robot. Autom. Lett.*, vol. 8, no. 5, pp. 2828–2835, May 2023, doi: 10.1109/LRA.2022.3221336.

[5] W. Wang, C. Wang, J. Liu, X. Su, B. Luo, and C. Zhang, "HVL-SLAM: Hybrid vision and LiDAR fusion for SLAM," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, art. no. 5706514, pp. 1–14, 2024, doi: 10.1109/TGRS.2024.3432336.

[6] Y. Liu, T. Jia, S. Guo, Y. Liu, K. Zhang, and Z. Du, "A robust and accurate stereo SLAM based on learned feature extraction and matching," *IEEE Trans. Instrum. Meas.*, vol. 74, art. no. 8514012, pp. 1–12, 2025, doi: 10.1109/TIM.2025.3614846.

[7] S. Zhang, J. Wu, K. Wang, and S. Deng, "TSO-BoW: Accurate long-term loop closure detection with constant query time via online bag of words and trajectory segmentation," *IEEE Robot. Autom. Lett.*, vol. 10, no. 5, pp. 4388–4395, May 2025, doi: 10.1109/LRA.2025.3550799.

[8] Z. Zhong, S. Chen, Q. Fu, J. Wang, and W. He, "FW-ORB-SLAM: A monocular visual SLAM algorithm for flapping-wing flying robots," *IEEE Robot. Autom. Lett.*, vol. 10, no. 12, pp. 13296–13303, Dec. 2025, doi: 10.1109/LRA.2025.3629994.

[9] J. Lee, M. Back, S. Hwang, and I. Chun, "Improved real-time monocular SLAM using semantic segmentation on selective frames," *IEEE Trans. Intelligent Transportation Systems*, vol. 24, no. 3, pp. 2800–2813, Mar. 2023, doi: 10.1109/TITS.2022.3228525.

[10] A. Cramariuc, L. Bernreiter, F. Tschopp, M. Fehr, V. Reijgwart, J. Nieto, R. Siegwart, and C. Cadena, "maplab 2.0 – a modular and multi-modal mapping framework," *IEEE Robot. Autom. Lett.*, vol. 8, no. 2, pp. 520–527, 2023, doi: 10.1109/LRA.2022.3227865.

[11] I. El Bouazzaoui, S. A. R. Florez, and A. El Ouardi, "Enhancing RGB-D SLAM performances considering sensor specifications for indoor localization," *IEEE Sensors J.*, vol. 22, no. 6, pp. 4970–4977, Mar. 2022, doi: 10.1109/JSEN.2021.3073676.

[12] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artificial Intelligence*, vol. 2012, article ID 376804, 19 pages, Nov. 2012, doi: 10.5402/2012/376804.

[13] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive review of k-means clustering algorithms," *International Journal of Advances in Scientific Research and Engineering*, vol. 7, no. 8, pp. 64–69, Aug. 2021, doi: 10.31695/IJASRE.2021.34050.

[14] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 224–236.

[15] S. M. Omohundro, "Five balltree construction algorithms," ICSI Technical Report TR-89-063, Dec. 1989.

[16] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 1177–1178. ACM, 2010.

[17] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.

[18] S. Garg, N. Sünderhauf, and M. Milford. Semantics for robotic mapping, perception and interaction: A survey. *Foundations and Trends in Robotics*, 8(1-2):1–224, 2021.

[19] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.