Hierarchical Unsupervised Topological SLAM

Ayush Sharma*[1], Yash Mehan*[1], Pradyumna Dasu[1], Sourav Garg[2], K. Madhava Krishna[1]

Fig. 1: [Left] A top-down view of Matterport3D [1] scene, [Right] An embodied agent's traversal and sequences obtained from trajectory unsupervised segmentation.

***Abstract*— In this paper we present a novel framework for unsupervised topological clustering resulting in improved loop detection and closure for SLAM. A navigating mobile robot clusters its traversal into visually similar topologies where each cluster (topology) contains a set of similar looking images typically observed from spatially adjacent locations. Each such set of spatially adjacent and visually similar grouping of images constitutes a topology obtained without any supervision. We formulate a hierarchical loop discovery strategy that first detects loops at the level of topologies and subsequently at the level of images between the looped topologies. We show over a number of traversals across different Habitat environments that such a hierarchical pipeline significantly improves SOTA image based loop detection and closure methods.**

Further, as a consequence of improved loop detection, we enhance the loop closure and backend SLAM performance. Such a rendering of a traversal into topological segments is beneficial for downstream tasks such as navigation that can now build a topological graph where spatially adjacent topological clusters are connected by an edge and navigate over such topological graphs.

I. INTRODUCTION

The SLAM problem has been widely studied in robotic and computer vision communities exploring various aspects of the problem. [2], [3] discuss about early SLAM frameworks. The SLAM taxonomy includes classification based on sensing modality (monocular SLAM [4], [5], those that incorporate depth data [6], and those that use LIDAR [7]), incorporate multiple robots [8], data-driven paradigms [9], and those that incorporate semantics and objects [10]. However, there have been very few approaches that integrate a topological understanding [11].

Nonetheless, topological understanding is beneficial to both localization and mapping, as amply demonstrated in [12]. In this paper, we exemplify a novel unsupervised topological SLAM framework that segments the observations (images) accrued during a traversal into clusters. Each such cluster of images demarcates a topology (see Figures 1, 3) and can now be represented with a single representative embedding. The images that constitute a cluster are also obtained from spatially and temporally adjacent locations, and hence we call a topological embedding also as a sequence embedding.

As a direct consequence of segmenting a traversal into such sequence embeddings, we demonstrate improved loop detection and closures, especially so for sequences viewing the same scene but from a very different and disparate approach direction (see Figure 3).

Specifically, the paper makes the following contributions:

1) Proposes a novel, and one of the first such, framework for unsupervised hierarchical topological SLAM.
2) We show enhanced loop detection and closure, exploiting the advantages offered by the hierarchical represen-

* Denotes Equal Contribution
[1] IIIT Hyderabad mkrishna@iiit.ac.in
[2] University of Adelaide sourav.garg@adelaide.edu.au
The authors thank MathWorks for their generous financial support.

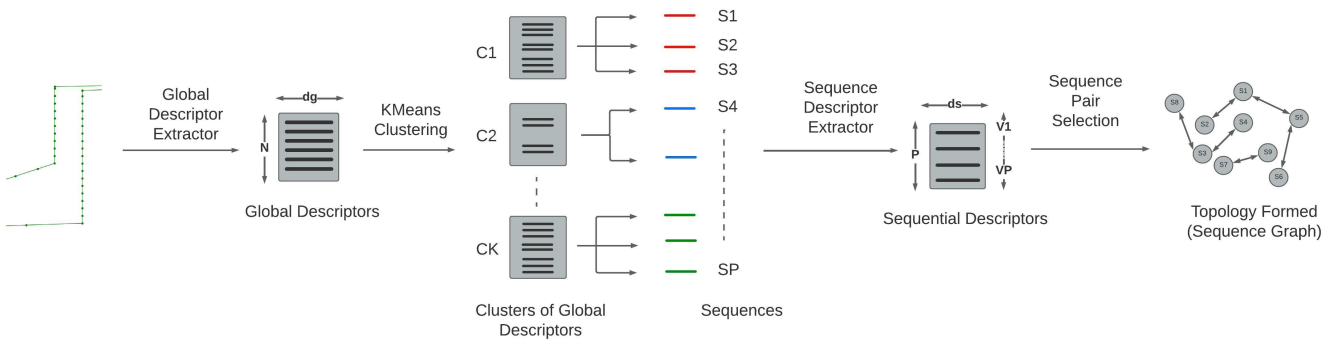


Fig. 2: Topology Formation Pipeline. Feature extraction, clustering then sequential descriptor extraction followed by pairing sequential descriptor based on their similarity score to output the topology formed(visualised as sequence graph). The sequences S_i's will be having variable length i.e. number of frame will be different in different frames. The edges between two nodes in sequence graph, says that both node (sequences) belong to same topology. Here, N and d_g denote total number of pose/frames and global descriptor dimension respectively. And, P, V and d_s denote total number of sequences, sequence descriptor and it's dimension respectively.

tation. We illustrate loop detection of sequences that observe the same topology or scene from disparate approach directions, making use of layered representations made possible due to the hierarchy.

3) Quantitatively, we portray consistently higher Precision-Recall values vis-a-vis baselines [13], [14] that do not resort to sequential descriptor matching for loop detection.
4) Further, we integrate our unsupervised topological SLAM framework with the popular RTABMAP to show improved loop detection and backend pose-graph optimization.

II. RELATED WORK

A. Image Representation

Architectures like DBoW2 [15], GeM [14], NetVLAD [16], and more recently, CosPlace [17] have been effectively demonstrated for global image representations for retrieval tasks. DBoW2 performs classical indexing and converts images into a bag-of-words representation, while building a hierarchical tree for approximating nearest neighbours in the image feature space and creating a visual vocabulary. Several recent works had kept focus on deep network methodologies [13], [14], [16], especially improving pooling step for better utilisation of visual information present in an image.

Recent work [18] shows that *K-STD* achieves better performance than the raw and standardized descriptors over the evaluated range of *K*. Comparatively better result by *K-STD* can be argued based on the exploitation of the clustering ability of global descriptors into meaningful clusters, i.e. visually similar images in one cluster and non-similar in separate clusters. Therefore, one default inherited property of global descriptors is to provide us better clustering (or grouping) of images, via any unsupervised clustering algorithms like KMeans and KMeans++.

B. Sequence Representation

Sequence-based place recognition has been extensively studied in the field of localization, with earliest approaches based on the post processing of a distance matrix computed by matching single image descriptor [19]–[23]. A recent trend in this area has been to use sequential *descriptors* to match places across reference and query traverses. *Facil et al.* [12], [24] used three basic techniques: concatenation of single image descriptors, fusion of the frame-level features with an FC layer, and integration over time of the single-image features via an LSTM network. Delta descriptors proposed in [25] provide an unsupervised method for sequential representation. SeqNet [26] proposed hierarchical sequential VPR which uses sequential descriptor based matching to guide single frame-based score aggregation. A more recent work [27] proposes the categorization of architectures depending on the stage where the fusion mechanism is applied, i.e. late fusion like GeM+CAT, early fusion like Timesformer [28] and intermediate fusion like SeqNet [26] and SeqVLAD [26], [27]. While sequence based approaches to VPR problem are generally superior to single image based approaches, use of sequential descriptors in particular reduces the cost of matching by incorporating temporal clues into the descriptor. This is because sequential descriptor methods summarize each sequence with a compact single vector and then perform the similarity search directly sequence-by-sequence.

However, a key issue persists in these sequential descriptor techniques: they are all demonstrated to be only effective for a *fixed* sequence length setting. This is preset for each training run for both SeqNet [26] and SeqVLAD [27], thus requiring a new fixed sequence length model to be trained every time. While the original SeqVLAD was not demonstrated for a variable length aggregation, our experiments show that it can be effectively used for sequence length unaware topology extraction and loop

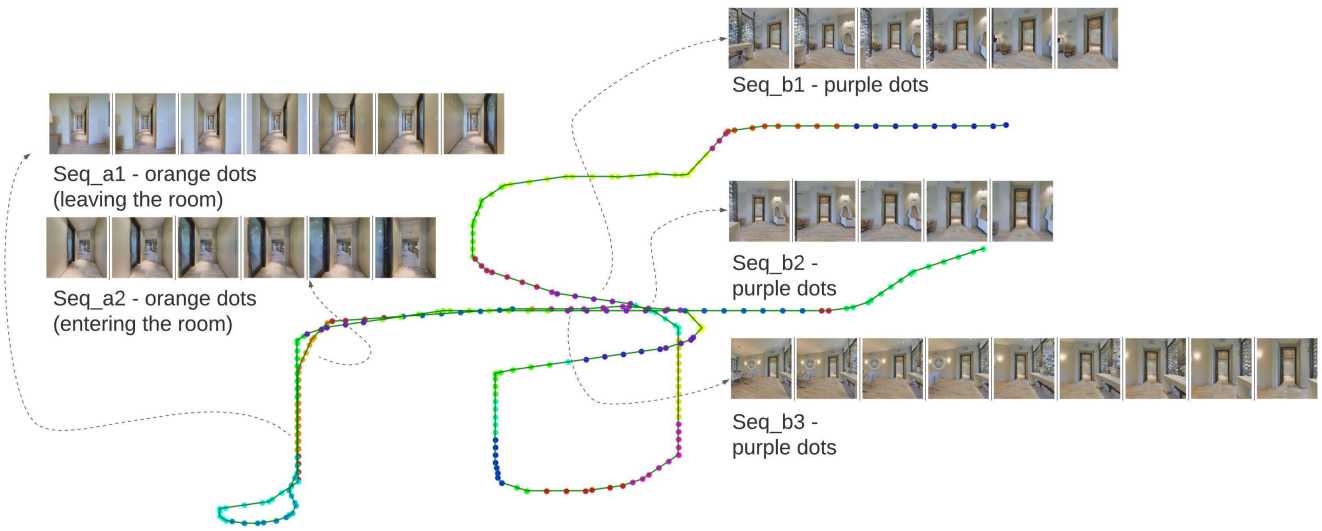


Fig. 3: Topology formation from agent's run. Figure visualises the segmentation of a traversal into sequences. Dots of a colour belong to one cluster. Sequences Seq_{a1} and Seq_{a2} belong to same cluster. So does Seq_{b1}, Seq_{b2} and Seq_{b3}.

closures, refer IV-A for the same.

III. METHODOLOGY

We first introduce the proposed pipeline of topology extraction. This is based first on clustering of image-level embeddings, followed by enforcing temporal connectivity on said clusters to obtain 'sequences' from within those clusters. Then, we discuss a pose-based topology estimation and demonstrate our pipeline for the task of loop detection and closure for visual SLAM.

A. Topology Formation

By topology we imply a collection of images whose individual embeddings are closer to the representative embedding of the entire collection vis-a-vis representative embeddings of other such collection. We also ensure that such collection of images are spatially close, as well as comprise of images viewing the same place from several different viewpoints, gathered potentially through multiple visits of that place during the agent's traversal. By the task definition, some of the observations would be temporally close, that is, consecutive image frames during the traversal. Our proposed topology formation (see Figure 2) is comprised of three steps: single image feature extraction, feature clustering to obtain sequences based on intra-cluster temporal connectivity, and finally, extraction of sequential descriptors which are the representative feature vectors or embeddings for that topology.

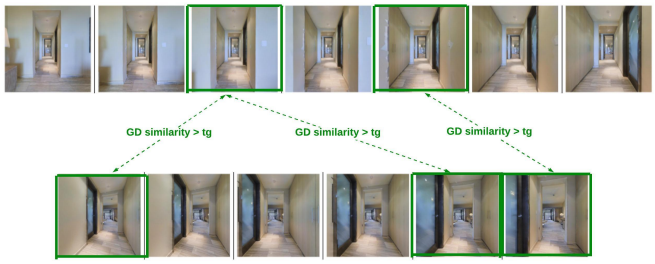
1) Feature Extraction: Given RGB images from the agent's traversal of the environment, say I_i $(i \in [1, N])$, we extract global descriptors D_i for each image. Our method is agnostic to the choice of the global descriptor, thus both classical methods like DBoW2 [15] or more recent deep network based architectures like NetVLAD [16] and GeM [13], [14] can be used. We used the latter due to their demonstrated high performance; specifically, we used ResNet101+GeM [13], [14] for obtaining a 2048-dimensional compact vector embedding as a global descriptor for each image.

2) Clustering: In this step, the extracted global descriptors are clustered using KMeans algorithm. We used the elbow curve technique to decide a suitable *K* (i.e., the number of centroids). This results in clusters with images that view the same place. These images would typically occur in the form of sequences, including those farther in timestamps due to multiple visits of the same place. At the same time, a cluster might comprise 'outliers', that is, it could contain images that belong to a different place which is physically far apart from the majority cluster members. These *K* clusters can be sub-divided into sequences, defined as a set of frames that are consecutive (that is, temporally connected).

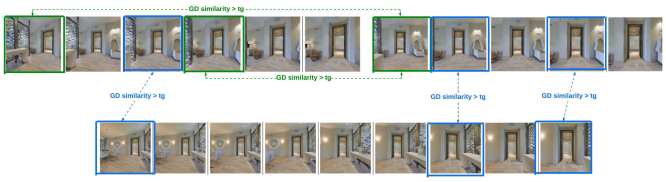
After clustering the global descriptors D_i, we obtained clusters C_k $(k \in [1, K])$. We sub-divide each of these clusters to obtain sequences S_{jk} $(j \in [1, P_k])$, where P_k refers to number of sequences in k^{th} cluster and each such sequence will belong to exactly one of the K clusters (see Figures 2, 3). Also, note that the number of frames will vary across sequences S_{jk} $(j \in [1, P_k])$, i.e. sequences of different lengths will be there.

3) Sequential Descriptor: From the sequences obtained from the previous step, S_{jk} for each cluster k, we compute their corresponding sequential descriptors V_{jk}. This is achieved through SeqVLAD [27], which we slightly adapt to obtain embeddings for *variable* length sequences, even though it is trained on fixed length sequences as described in section II-B.

While we obtained sequences as members of the clusters formed through single images, a more informed representation of a place is achieved through the sequential descriptors. Thus, we can now compare sequential descriptors *both* within and across clusters to finally decide the topologies. For



(a) Image matching between Seq_{a1} and Seq_{a2}



(b) Image matching between Seq_{b1}, Seq_{b2} and Seq_{b3}

Fig. 4: Intra-Topology Matching. exhaustive image to image matching between selected sequence pair from the sequence graph to output the loop pairs is shown. Green pairs show loop pairs with visible overlap. Blue pairs show loop pairs having seemingly less visual overlap.

this purpose, we use cosine similarity $s_{pp'} \in [-1, 1]$ to compare two sequence descriptors V_p and $V_{p'}$ where $p, p' \in [1, \sum_k P_k]$ refer to the indices of sequences regardless of their cluster k. If $s_{pp'} \geq t_s$, then sequences S_p and $S_{p'}$ belongs to the same topology.

Visualisation of the topology so formed can be done in graphical way, where nodes represent sequences and an edge between two nodes implies that they belong to the same topology. We can refer to the obtained graph as a sequence graph or sequence based topological graph (see Figure 2).

B. Intra-Topology Matching

Following Section III-A, one can get a sequence based topological graph (sequence graph) for an agent's traversal in an environment. Then, intra topological matching can be done as follows: for all the sequence pairs having an edge between them in the sequence graph, an exhaustive image to image (first sequence S_{jk} to second sequence $S_{j'k}$) global descriptor (D_i) matching is performed. The image pairs having a global descriptor similarity score above the threshold t_g are detected as loop pairs (see Figure 4).

C. SLAM Pipeline

Formally, given a visual traversal, the loop detection task is to compute correct loop pairs within a radius, say 5 m. Vision based loop detection task utilises VPR (visual place recognition) [24] solutions as core modules. The topology formation technique along with the intra-topological matching forms a two-stage hierarchical loop detection pipeline, which detects loop closure candidates at the sequence and the image levels respectively. We propose a SLAM pipeline whose frontend comprises of this hierarchical loop detection pipeline and a robust local feature matching pipeline like OriNet [29]–[31], with the backend optimisation handled by G2O [32]. One widely accepted SLAM framework, RTABMAP [33], performs the same task but employs DBoW2 based loop detection system with additional loop hypothesis filtering modules before local feature matching step.

Later, we show that augmenting RTABMAP with our proposed hierarchical system's loop pairs provides a better final trajectory optimisation with G2O [32] (Section IV-B).

IV. EXPERIMENTATION AND RESULTS

A. Setup

The dataset used was Matterport3D [1] and Habitat-sim [34] for simulating an embodied agent traversing the environment. 12 environments with 2 trajectories each were chosen at random for training of the sequence descriptor extractor. The method of **III A2** was employed to extract sequences for each trajectory. The obtained training sequences, culled to a static length(5 for experiments in this paper) were employed during training SeqVLAD and one of late fusion model [27] ResNet18l3+GeM+CAT. Small sequence length is in response to fast visual change in indoor setting as compared to outdoor scenarios. Finally, our models have been trained following same training method, as defined in [27].

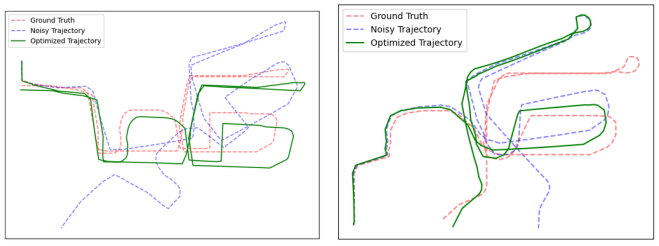
During inference, given the agent's traversal, method as in **III A2** was employed to extract variable length sequence. SeqVLAD is an intermediate fusion model [27] and extended version of NetVLAD [16]. VLAD [16], [27] layer output is independent of height, width and depth dimensions of the feature maps obtained after convolution layers, but not of the sequence dimension. Keeping this in mind, we infer one sequence at a time through SeqVLAD i.e. inference batch size one that enabled us to extract fixed length representations for variable length sequences.

B. Augmenting RTABMAP

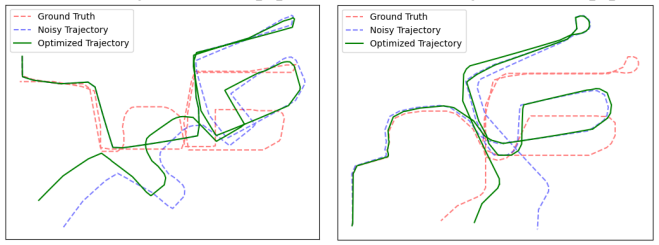
Unsupervised SLAM indoors is a challenging yet promising avenue due to presence of repeating scenes and potential of opposite view loop pairs while re-traversing paths. The ability of our system to capture loop pairs having extreme viewpoints difference has been highlighted in figure 6, which RTABMAP [33] and DBoW2 fails to detect.

Post detection of opposite view pairs by our proposed method, OriNet [29]–[31], accepted for robustness to aggressive viewpoint changes, is employed for feature matching. Subsequently the frame transform between the loop pairs is passed to the backend optimizers. Figure 5 exhibits enormous boost in backend optimization of the trajectory post augmentation of loop pairs. The same can be seen in Table I wherein the mean Absolute Pose Error of augmented loop pairs is

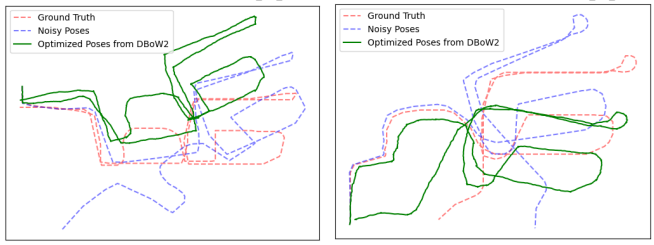
significantly better than RTABMAP's native or DBoW2's detected loops.



(a) Run 1 Augmented loop pairs (b) Run 2 Augmented loop pairs



(c) Run 1 RTABMAP loop pairs (d) Run 2 RTABMAP loop pairs



(e) Run 1 DBoW2 loop pairs (f) Run 2 DBoW2 loop pairs

Fig. 5: Pose-graph optimisation results using G2O. We obtained the noisy trajectory by adding Gaussian noise to the ground truth odometry information. External noise has been added to account for drift accumulated in the real world dataset. The subfigures (a) and (b) compare optimization of noisy trajectories based on augmented loop pairs, (c) and (d) show optimization of noisy trajectories based on RTABMAP's native loop pairs only. (e) and (f) show optimisation based on DBoW2's loop pairs. RTABMAP + Our augmented loop pairs enhance the optimization as compared to RTABMAP's native or DBoW2's loop pairs. This observation is consistent across numerous possible noise having different mean and variance.

C. Performance Analysis

A typical global descriptor based loop detection system considers every frame as the query and all the previous frames (except a temporally adjacent ones) as the reference set. To filter the detected loop pairs, all retrieved images which have their descriptor similarity score at-least some threshold t_g is considered. We make use of ResNet101+GeM descriptors for baseline comparison with our proposed hierarchical system. In addition, we make baseline comparison with DBoW2 based loop detection system as a representative for BoW (bag of words) class of typical classical descriptor system.

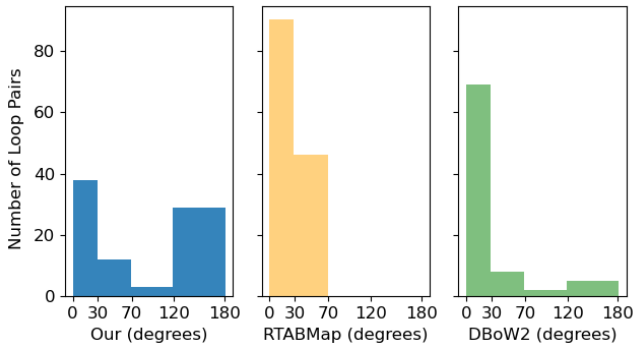


Fig. 6: Ground truth rotational difference between candidates of a loop pair in degrees for different loop detection pipeline. Detection of obtuse and opposite view loop pairs is significantly improved due to our formulation as compared to RTABMAP's and DBoW2's loop detection. This compensates any reduction observed in the proposed system while dealing with the loop pairs that have very close or similar viewing angle. The improved backend optimization due to the augmented framework confirms this in Figure 5 and Table I.

TABLE I: Absolute Pose Error (APE) of trajectories w.r.t ground truth trajectories.

APE w.r.t Ground Truth	**run1**			**run2**		
Trajectory	***min***	***mean***	***max***	***min***	***mean***	***max***
Noisy	0.011	1.715	6.457	0.224	0.930	3.576
Optimisd with DBoW2 loop pairs	0.084	0.956	2.475	0.127	0.782	1.77
Optimised with RTABMAP native loop pairs	0.020	1.289	5.190	0.189	**0.649**	1.689
Optimised with RTABMAP+Our loop pairs	**0.018**	**0.312**	**0.757**	**0.08**	0.69	**1.673**

1) Precision and Recall: For loop detection system, we define recall as percentage of correct predicted loop pairs by the system out of all possible correct loop pairs, and precision as percentage of correct predicted loop pairs out of total predicted loop pairs. PR curves (see Figure 7) show that the proposed method is typically better for every pair of precision-recall values as the respective thresholds for the methods are varied vis a vis the ResNet+GeM method. The proposed method is on-par with the BoW methods but specifically is able to recall better loop pairs with significant disparity in viewing angles (see Figure 6)

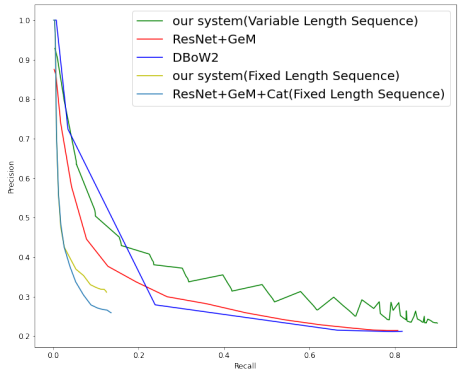


Fig. 7: PR-curves for loop pairs from different systems

2) PR curve for fixed sequence length based loop detection systems: In Figure 7, we showcase the advantages of SeqVLAD inherent ability to handle variable length sequences at inference time. When SeqVLAD is forced to use fixed length clusters at inference time the recall drops considerably whereas variable length SeqVLAD has a much better recall. Therefore, proposed hierarchical system faces significant decrement in the performance when constrained to fixed length sequences for the results in this paper.

V. CONCLUSION AND FUTURE WORK

The aim of this work is to show that a clever use of image level and sequence level information for multi-stage hierarchical loop detection augments SLAM frameworks' capability for detecting loop closures and showing significant improvement over deep global descriptor methods.

Our work detects loop closures between perspective images of the same place with a wide shift in the viewpoint, where the classical methods fail to keep up. This opens up new avenues for improving the pose-graph optimization in SLAM frameworks that utilize methods which rely on the agent revisiting a scene in similar viewpoints.

REFERENCES

[1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[2] B. Siciliano and O. Khatib, *Springer Handbook of Robotics*, 2008.

[3] D. Fox, S. Thrun, and W. Burgard, *Probabilistic Robotics*, 2005.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, pp. 1147–1163, 2015.

[5] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014.

[6] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 2011, pp. 127–136.

[7] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and R. Daniela, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5135–5142.

[8] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, pp. 2022–2038, 2021.

[9] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *Advances in neural information processing systems*, 2021.

[10] P. Parkhiya, R. Khawad, K. M. Jatavallabhula, B. Bhowmick, and K. M. Krishna, "Constructing category-specific models for monocular object-slam," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, 2018.

[11] S. S. Puligilla, S. Tourani, T. S. Vaidya, U. S. Parihar, R. K. Sarvadevabhatla, and K. M. Krishna, "Topological mapping for manhattan-like repetitive environments," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6268–6274, 2020.

[12] J. M. Fácil, D. Olid, L. Montesano, and J. Civera, "Condition-invariant multi-view place recognition," *ArXiv*, vol. abs/1902.09516, 2019.

[13] J. Revaud, J. Almazán, R. S. de Rezende, and C. R. de Souza, "Learning with average precision: Training image retrieval with a listwise loss," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5106–5115, 2019.

[14] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, pp. 237–254, 2016.

[15] 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, October 2012.

[16] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[17] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[18] S. Schubert, P. Neubert, and P. Protzel, "Unsupervised learning methods for visual place recognition in discretely and continuously changing environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2020.

[19] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 261–286, 2007.

[20] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1643–1649.

[21] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Are you able to perform a life-long visual topological localization?" *Autonomous Robots*, pp. 1–21, 2017.

[22] O. Vysotska and C. Stachniss, "Relocalization under substantial appearance changes using hashing," in *Proceedings of the IROS Workshop on Planning, Perception and Navigation for Intelligent Vehicles, Vancouver, BC, Canada*, vol. 24, 2017.

[23] P. Neubert, S. Schubert, and P. Protzel, "A neurologically inspired sequence processing model for mobile robot place recognition," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3200–3207, 2019.

[24] S. M. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. D. Cox, P. Corke, and M. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, pp. 1–19, 2016.

[25] S. Garg, B. Harwood, G. Anand, and M. Milford, "Delta descriptors: Change-based place representation for robust visual localization," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5120–5127, 2020.

[26] S. Garg and M. J. Milford, "Seqnet: Learning descriptors for sequence-based hierarchical place recognition," *IEEE Robotics and Automation Letters*, 2021.

[27] R. Mereu, G. Trivigno, G. Berton, C. Masone, and B. Caputo, "Learning sequential descriptors for sequence-based visual place recognition," 2022.

[28] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[29] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.net: Keypoint detection by handcrafted and learned cnn filters," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5835–5843, 2019.

[30] J. M. Dmytro Mishkin, Filip Radenovic, "Repeatability Is Not Enough: Learning Discriminative Affine Regions via Discriminability," in *Proceedings of ECCV*, Sept. 2018.

[31] F. R. J. M. Anastasiya Mishchuk, Dmytro Mishkin, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proceedings of NeurIPS*, Dec. 2017.

[32] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3607–3613.

[33] M. Labbé and F. Michaud, "Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[34] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.