Advances in Computer, Signals and Systems (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2026.100113
ISSN 2371-8838 Vol. 10 Num. 1

A Survey on Mobile Robot Visual Relocalization in Complex Dynamic Scenes

Shichu Sun[1,a,*]

[1]*Yunnan Normal University, Kunming, Yunnan, China*
[a]*2324100044@ynnu.edu.cn*
[*]*Corresponding author*

Keywords: Visual Relocalization, Mobile Robots, Dynamic Scenes, Robustness

Abstract: Visual relocalization is key to enabling continuous autonomous navigation of mobile robots. This survey systematically reviews research progress on visual relocalization technology in complex dynamic scenes, focusing on the issues of excessive computational load and mismatches caused by dynamic disturbances in real environments. It explores the evolutionary trajectory from global representations and local matching to semantic learning methods, revealing that the core mechanisms for handling dynamic disturbances have shifted from passive feature filtering to "active suppression" using spatial attention, adaptive matching, and soft weighting. Taking into account the deployment requirements of mobile robots, the paper summarizes engineering solutions that encompass lightweight matching and multimodal priors, and envisions future research directions such as deep semantic-geometric collaboration and incremental map evolution. These insights provide theoretical reference and practical guidance for the design of robust localization systems.

1. Introduction

During long-term autonomous operation of mobile robots, visual relocalization plays a crucial role in drift correction and as a failsafe mechanism. When conventional visual odometry or SLAM systems fail due to aggressive motion or severe occlusion, the relocalization module can recompute the camera's 6-DoF global pose from a single frame using an offline map [1, 2]. However, dynamic changes in the environment have always been the core bottleneck for large-scale deployment of this technology; high-frequency dynamic disturbances completely break the "static world" assumption on which traditional algorithms rely [3]. In real-world scenes, large-scale dynamic occlusions often cause a dramatic reduction in effective feature points. For example, in public datasets such as Cambridge Landmarks [2] that include dense crowds and vehicle occlusions, these disturbances not only significantly reduce localization success rates but also increase the relative trajectory error by over 20% compared to the static baseline, ultimately leading to tracking failure [4, 5]. To address the above challenges, the main contributions of this paper are threefold:

(1) A unified evaluation framework is constructed along three dimensions—short-term dynamics, periodic appearance changes, and long-term structural evolution—and focuses on three representative benchmarks: 7-Scenes, Cambridge Landmarks, and Indoor6.

(2) We systematically review and compare the performance and engineering trade-offs of methods

such as global retrieval, sparse feature matching, semantic augmentation, and multimodal fusion in dynamic scenes.

(3) We distill practical engineering recommendations and future research directions aimed at the deployment of mobile robot systems.

2. Problem Definition and Evaluation Benchmarks

The visual relocalization problem can be formally defined as follows: given the current query image and a pre-constructed offline map, solve for the camera's 6-DoF pose in the coordinate system of. In complex dynamic scenes, the image often contains dynamic objects (such as pedestrians and vehicles) not present in the map or significant appearance differences (e.g., lighting or seasonal changes). These observational inconsistencies are the fundamental causes of matching errors and pose drift [6]. To objectively evaluate visual relocalization algorithms under different types of complex scenes, this paper focuses on three representative and widely used benchmark datasets: 7-Scenes [1], Cambridge Landmarks [2], and Indoor6 [7], as shown in figure 1.

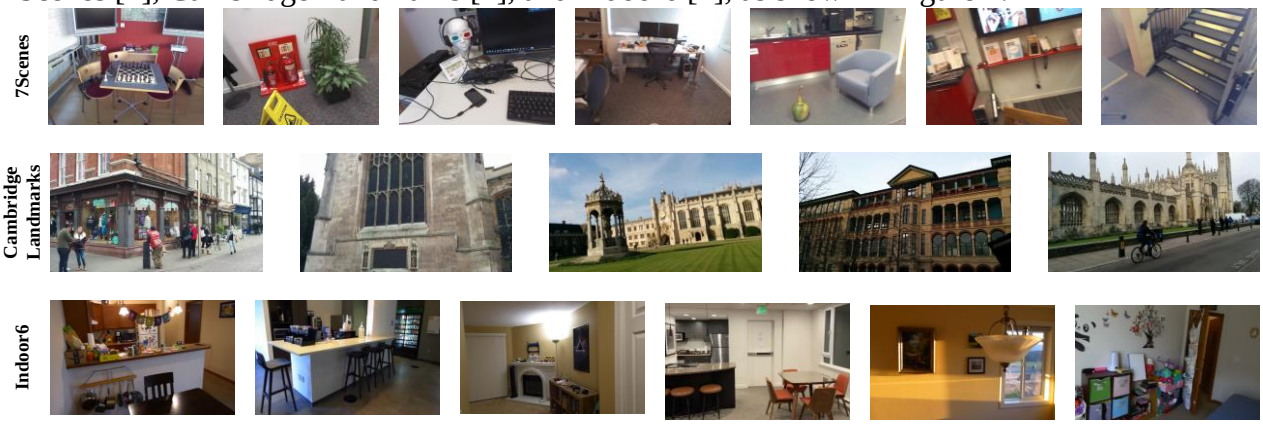

Figure 1: Example scenes from the 7-Scenes, Cambridge Landmarks, and Indoor6 datasets.

These three datasets each emphasize different aspects in terms of spatial scale, types of dynamics, and structural change cycles, collectively covering the common challenges encountered in practical deployment of mobile robots.

(1) 7-Scenes (indoor small-scale, short-term dynamics): This dataset includes motion blur, local object movement, and short-term occlusions in office and household scenes. It is commonly used to evaluate whether an algorithm can maintain a high localization success rate within a 5 cm/5 ° error threshold under short-term high-frequency dynamics and texture-scarce environments.

(2) Cambridge Landmarks (outdoor large-scale, periodic appearance changes): This dataset targets outdoor urban street scenes and exhibits pedestrian/vehicle occlusions and periodic appearance changes such as day-night cycles and seasonal variations. It is suitable for evaluating a method's long-term stability under large-scale retrieval and cross-time appearance drift (typically measured by the median translation and rotation error of the pose).

(3) Indoor6 (indoor large-scale, long-term structural changes): This dataset simulates long-term structural changes such as furniture displacement and interior rearrangement. It examines a method's generalization and map update capabilities when the geometric structure has undergone substantial changes.

Using these three benchmarks for comprehensive testing enables a thorough evaluation along the three dimensions of short-term dynamics, periodic appearance changes, and long-term structural evolution, while avoiding dispersing evaluation attention to too many minor datasets. This makes the conclusions more practically actionable and comparable.

3. Visual Relocalization Methods

With the evolution of visual localization technology towards deep learning [8], modern relocalization systems have gradually become hybrid coarse-to-fine architectures. We categorize them into three types: global methods for candidate frame selection, local geometric methods for precise pose estimation, and high-level learning methods that enhance system robustness.

3.1. Coarse Global Representation and Retrieval

Global representation methods convert the visual relocalization problem into an image retrieval problem. Arandjelović et al. [6] first proposed the NetVLAD network, which designs a generalized local aggregation descriptor (VLAD) layer and seamlessly integrates it into a convolutional neural network for end-to-end image place recognition. This enables efficient aggregation of the visual features of an entire image into a high-dimensional global descriptor. Subsequently, Sarlin et al. [9] proposed the large-scale hierarchical localization framework HLoc, which innovatively decouples global image retrieval from local feature matching. HLoc establishes the hybrid localization paradigm of "first roughly filter candidate frames with global features, then perform precise registration with local features". However, traditional global descriptors can easily fail in dynamic scenes. Experiments by Germain et al. [10] show that when dynamic elements (such as buses or crowds) occupy more than 50% of the pixels, the image's global feature distribution is severely disrupted, leading to a drastic drop in matching accuracy for large-scale scenes. To overcome this deficiency, Luo et al. [11] proposed a scalable visual localization scheme that introduces a lightweight spatial attention mechanism in the global retrieval stage. This mechanism guides the network during training to adaptively suppress pixels from highly dynamic regions and force focus on long-term static structures such as building contours. Compared to a traditional NetVLAD baseline network, the model with the attention filtering mechanism achieves a significant improvement of over 10% in Top-1 retrieval recall under complex occlusion conditions, while keeping computational cost roughly the same. This effectively mitigates the disturbance of the descriptor distribution caused by dynamic scenes.

3.2. Sparse Local Features and Geometric Registration

The core of local feature matching lies in establishing robust 2D–3D correspondences and then using a PnP (Perspective-n-Point) algorithm to solve for the precise pose. The SuperGlue method proposed by Sarlin et al. [12] is a breakthrough in this field. It introduces a graph neural network (GNN) based on self-attention and cross-attention mechanisms, jointly reasoning about contextual feature information between image pairs and solving a differentiable optimal transport problem. This significantly improves the number of inlier matches and the accuracy of geometric registration. However, in dynamic indoor scenes with many walking people, Zhao et al. [13] point out that extracting local features on moving objects can easily produce false correspondences across views. These mismatches greatly increase the computational burden of subsequent RANSAC-based geometric verification and can even lead to failure of the pose solution. Moreover, the dense attention-based graph interactions that SuperGlue relies on incur substantial computational overhead: even on high-end desktop GPUs, inference can take on the order of 70–100 ms [13]. This makes it difficult to deploy SuperGlue in real time on resource-constrained mobile robot platforms. To address the tradeoff between real-time performance and robustness, Lindenberger et al. [14] proposed LightGlue, an efficient feature matching network. This approach introduces an adaptive computational stopping mechanism on top of the graph neural network, which dynamically adjusts the number of inference layers and depth based on the visual overlap and matching difficulty of the input image pair.

Compared to the classic SuperGlue dense matching [13], employing this adaptive strategy can reduce the overall matching latency by about 30%, while still maintaining over 90% localization success rate under low frame-rate input conditions [14, 15]. Therefore, sparse adaptive matching is often considered an effective solution for edge deployment on mobile robots in practice.

3.3. Semantic and End-to-End Learning Methods

To avoid the complex feature extraction and verification steps in traditional geometric matching, deep learning has opened a new direction of direct pose regression. Kendall et al. [2]'s PoseNet pioneered an end-to-end pose regression framework. It uses a deep convolutional neural network to directly regress the camera's 6-DoF pose from a single RGB image. This achieves extremely fast single-frame inference, but its translation error in large-scale scenes often reaches around 2 m. To improve indoor localization accuracy, Brachmann et al. [16] proposed the DSAC algorithm based on scene coordinate regression. It innovatively transforms the traditional RANSAC hypothesis selection process into a probabilistic differentiable form, enabling end-to-end joint training of the neural network and geometric optimization. More recently, the same team proposed the ACE (Accelerated Coordinate Encoding) framework [17], which uses high-speed optimization of a multi-layer perceptron (MLP) to implicitly store the scene environment. ACE can complete scene mapping within a few minutes and, using an extremely lightweight network model of about 4 MB, achieves over 90% relocalization success rate on the 7-Scenes dataset. Furthermore, in the area of multimodal extension of end-to-end regression, Yang et al. [18] proposed the LiSA framework, which distills semantic-aware knowledge into a LiDAR-based scene coordinate regression network. This method effectively suppresses the negative interference of dynamic objects and repetitive physical structures on coordinate prediction, significantly enhancing localization robustness in complex environments without increasing inference parameters. For explicitly handling dynamic disturbances, actively filtering using a semantic segmentation network is currently the most direct approach. Abati et al. [19]'s Panoptic-SLAM seamlessly integrates a panoptic segmentation model into the state estimation pipeline. Through panoptic segmentation, it explicitly filters out potential dynamic instances, removing interference at the observation source. Its absolute trajectory error (ATE) is improved by a factor of four compared to visual odometry methods without semantics. However, pure hard semantic removal has fragile failure boundaries. Liu et al. [20] point out that in dense crowds, aggressively removing person pixels can lead to a loss of over 80% of feature points in the local view. Once the remaining matched points do not meet the minimum requirements for geometric solving, localization will completely fail. To address the problem of feature depletion caused by over-removal, Soares et al. [4] proposed an efficient filtering mechanism for dynamic and long-term changing environments. This algorithm not only relies on object detection, but also precisely identifies and retains static keypoints within dynamic bounding boxes, effectively avoiding blind "one-size-fits-all" removal. Additionally, Yan et al. [21] proposed a long-term visual localization solution aided by mobile sensors. This method employs probabilistic "soft semantic weights" combined with an IMU prior for multimodal fusion. By using inertial sensor priors to provide effective search boundary constraints for visual solving, it has been proven a reliable strategy to prevent relocalization divergence in highly dynamic scenes [19, 22]. It should be noted, however, that semantic-based methods are highly dependent on the generalization capability of the front-end segmentation model to unseen scenes; their overall stability under cross-domain deployment still needs further verification.

4. Deployment Suggestions and Prospects

This survey reviews the development of mobile robot visual relocalization techniques in complex dynamic scenes, revealing that current research is gradually shifting from passive removal of dynamic

regions toward active modeling that understands scene structure. Focusing on computational constraints and environmental uncertainties in real robotic deployments, we offer the following engineering guidance recommendations:

(1) Adaptive matching and coarse-to-fine framework design. On computation-limited mobile robot platforms, we recommend a hierarchical strategy of "global retrieval – local precise matching". Use global features to filter candidate poses, and then apply sparse adaptive matching for fine geometric estimation. This approach maintains real-time performance while preserving localization accuracy.

(2) Avoid overly aggressive semantic hard-removal strategies. In scenes with weak texture or sparse features, completely relying on semantic segmentation to hard-remove dynamic regions may lead to insufficient geometric constraints. In contrast, introducing a soft-weighting mechanism to retain low-confidence static edges can improve the stability of pose estimation.

(3) Fusion of physical priors and multimodal constraints. For complex environments with frequent dynamic occlusions, prioritize using long-term stable structures (such as building outlines) to construct geometric constraints, and combine inertial sensors (e.g. IMUs) for short-term motion compensation. This enhances system robustness.

Furthermore, given the current limitations of visual relocalization in complex environments, future research will focus on strengthening the collaborative modeling of semantic information and geometric structure. This aims to improve the model's 3D reasoning of stable scene elements under dynamic occlusions and to validate its cross-scene generalization performance in larger and more diverse environments. Additionally, for long-term operational needs, incremental map evolution and local update mechanisms will be key directions. By adaptively optimizing changing areas, these approaches can reduce map maintenance costs and push visual relocalization systems toward truly long-term stable autonomous operation.

References

[1] Shotton, Jamie, Glocker, Ben, Zach, Christopher, Izadi, Shahram, Criminisi, Antonio, and Fitzgibbon, Andrew. "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2930–2937, 2013. doi: 10.1109/CVPR.2013.377

[2] Kendall, A., Grimes, M., and Cipolla, R. "PoseNet: A convolutional network for real-time 6-DOF camera relocalization". Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2912–2920, 2015. doi: 10.1109/ICCV.2015.333

[3] Kendall, A., Gal, Y., and Cipolla, R. "Modelling Uncertainty in Deep Learning for Camera Relocalization". arXiv preprint arXiv:1509.05909v2, 2016.

[4] Soares, J. C. V., and others. "Visual Localization and Mapping in Dynamic and Changing Environments". Journal of Intelligent & Robotic Systems, 109:95, 2023. doi: 10.1007/s10846-023-02019-6

[5] Toft, C., and others. "Long-Term Visual Localization Revisited". IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4):2074–2088, 2020. doi: 10.1109/TPAMI.2020.3032010

[6] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. "NetVLAD: CNN architecture for weakly supervised place recognition". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5297–5307, 2016. doi: 10.1109/CVPR.2016.572

[7] Do, T., and Sinha, S. "Improved Scene Landmark Detection for Camera Localization". Proceedings of the International Conference on 3D Vision (3DV), 975–984, 2024. doi: 10.1109/3DV62453.2024.00069

[8] Huang, K. "Overview of Visual SLAM Technology: From Traditional to Deep Learning Methods". Advances in Computer, Signals and Systems, 7(10):76–81, 2023. doi: 10.23977/acss.2023.071011

[9] Sarlin, Paul-Edouard, Cadena, Cesar, Siegwart, Roland, and Dymczyk, Martin. "From coarse to fine: Robust hierarchical localization at large scale". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12716–12725, 2019. doi: 10.1109/CVPR.2019.01300

[10] Germain, H., and others. "Learned place recognition". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. doi: 10.1109/CVPR52729.2023.00890

[11] Luo, Z., and others. "Scalable visual localization". Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2024. doi: 10.1109/ICRA.2024.101247

[12] Sarlin, Paul-Edouard, DeTone, Daniel, Malisiewicz, Tomasz, and Rabinovich, Andrew. "SuperGlue: Learning

feature matching with graph neural networks". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4938–4947, 2020. doi: 10.1109/CVPR42600.2020.00499
[13] Zhao, X., and others. "Robust feature matching". Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. doi: 10.1109/ICCV.2023.00714
[14] Lindenberger, P., Sarlin, P.-E., and Pollefeys, M. "LightGlue: Local Feature Matching at Light Speed". Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. doi: 10.1109/ICCV51070.2023.01633
[15] Chen, Y., and others. "Geometry-aware localization". IEEE Transactions on Robotics, 2024. doi: 10.1109/TRO.2024.3351982
[16] Brachmann, Eric, Krull, Alexander, Nowozin, Sebastian, and others. "DSAC – Differentiable RANSAC for Camera Localization". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6692–6701, 2017. doi: 10.1109/CVPR.2017.265
[17] Brachmann, E., Cavallari, T., and Prisacariu, V. "Accelerated Coordinate Encoding: Learning to Relocalize in Minutes using RGB and Poses". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2736–2746, 2023. doi: 10.1109/CVPR52729.2023.00267
[18] Yang, B., and others. "LiSA: LiDAR Localization with Semantic Awareness". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. doi: 10.1109/CVPR52733.2024.01123
[19] Abati, G. F., and others. "Panoptic-SLAM: Visual SLAM in Dynamic Environments using Panoptic Segmentation". Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2024. doi: 10.1109/ICRA50011.2024.10611425
[20] Liu, S., and others. "Semantic-geometry fusion". arXiv preprint arXiv:2501.09876, 2025.
[21] Yan, S., and others. "Long-term Visual Localization with Mobile Sensors". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. doi: 10.1109/CVPR52729.2023.01639
[22] Wang, F., and others. "GLACE: Global Local Accelerated Coordinate Encoding". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. doi: 10.1109/CVPR52733.2024.00492