Pose360: Metric-Scale Visual Odometry by Grounding Learned Features with LiDAR depth

Kemal Mudie Tosora[1,2], Seher Kanwal[1,2], Seung-Ik Lee[1,2*]
[1]Department of Artificial Intelligence, University of Science & Technology, Daejeon, South Korea
[2]Electronics & Telecommunications Research Institute, Daejeon, South Korea
kemal.tosora@etri.re.kr, seher@etri.re.kr, the_silee@etri.re.kr

***Abstract*—Metric-scale state estimation is a cornerstone for autonomous systems, yet single-modality solutions often fail in challenging real-world environments. Visual odometry suffers from inherent scale ambiguity, while LiDAR odometry can be fragile in geometrically sparse scenes. To overcome these limitations, we propose Pose360, a novel Visual-LiDAR Odometry (V-LIO) system that robustly fuses a 360° panoramic camera and a 360° LiDAR. Our approach leverages learned features, SuperPoint and LightGlue, to establish strong 2D visual correspondences, which are then lifted into a sparse set of metric 3D-3D point correspondences using depth information from a synchronized LiDAR point cloud. The 6-DoF relative pose is then computed efficiently via a closed-form SVD-based solution. This focused fusion strategy directly resolves visual scale ambiguity without requiring complex non-linear optimization. We perform a rigorous quantitative evaluation on a challenging $450\,\mathrm{m}$ real-world dataset, demonstrating that our system achieves high global consistency with a translation Absolute Trajectory Error (ATE) of just $0.177\,\mathrm{m}$ RMSE against a high-fidelity LiDAR SLAM ground truth. We further validate its real-world applicability by successfully integrating Lidar360Pose as the core odometry engine in a full robotic navigation stack, proving it is an accurate and reliable solution for metric state estimation.**

***Keywords*—Visual Odometry, Sensor Fusion, Pose Estimation, LiDAR, Panoramic Camera, Autonomous Systems**

I. Introduction

Accurate 6-DoF (Degree of Freedom) pose estimation is a fundamental capability for intelligent systems, enabling tasks from robot navigation to augmented reality [1], [2]. While vision-based methods using monocular cameras are highly proficient at tracking features in textured environments [3], they suffer from a critical, inherent limitation: scale ambiguity. Without an external reference, a purely visual system cannot distinguish a small motion near a close object from a large motion near a distant one, preventing the recovery of a true metric trajectory [4]. Depth sensors like LiDAR resolve this ambiguity by providing direct metric measurements. However, traditional LiDAR odometry, which relies on scan-to-scan registration [5], can be fragile in geometrically self-similar or sparse environments like long corridors.

This paper explores a fusion strategy that combines the strengths of both modalities. We propose that instead of relying on direct LiDAR scan registration for motion estimation, a more robust approach is to use LiDAR data for the sole purpose of grounding strong 2D visual feature matches in metric 3D space. State-of-the-art learned feature matchers like SuperGlue [6] provide a robust front-end for finding reliable correspondences. Our core idea is to leverage such feature matchers to resolve the inherent scale ambiguity by augmenting these 2D matches with precise depth from a synchronized LiDAR.

To this end, we introduce **Pose360**, a system that implements this focused fusion strategy. Our pipeline begins by establishing a sparse set of high-confidence 2D correspondences between two 360° panoramic images. We then project the LiDAR point cloud into the image coordinate system to create an efficient depth lookup structure. Each 2D feature match is lifted into a pair of 3D-3D correspondences in metric space. This formulation transforms the problem from a complex, non-linear visual optimization into a well-posed absolute orientation problem that can be solved efficiently in closed form.

Our contributions are:

- **A Metric-Grounded Visual Odometry Framework:** We propose and implement a novel Pose360 pipeline that directly resolves the scale ambiguity of a learned visual feature front-end by augmenting 2D matches with precise LiDAR depth.
- **Real-World Robotic Evaluation:** We provide a rigorous quantitative analysis of our system's trajectory accuracy on a mobile robot, validating its performance against a high-fidelity LiDAR SLAM ground truth in a large-scale, complex environment.
- **System-Level Validation:** We demonstrate the practical efficiency and robustness of our method by integrating it as the core local odometry engine within the Visual Teach-and-Replay navigation framework.

II. Related Work

Our research is situated at the confluence of visual state estimation, LiDAR-based mapping, and multi-sensor fusion. We review key developments in these areas to contextualize our contribution.

A. Visual Odometry and the Challenge of Scale

Visual Odometry (VO) estimates camera motion by tracking features across images. Landmark-based approaches rely on sparse, hand-crafted features like SIFT [7] or ORB [8], which

*Corresponding author

form the backbone of highly successful SLAM systems such as ORB-SLAM3 [3]. The advent of learned local features, such as SuperPoint [9], and graph-based matchers like SuperGlue [6], has significantly improved the robustness of the data association front-end. Our work deliberately leverages these state-of-the-art learned components for their superior matching performance.

However, a fundamental limitation of all monocular VO systems is their inability to observe metric scale [4]. The estimated trajectory is only known up to an arbitrary scale factor, rendering it insufficient for many robotic tasks. While Visual-Inertial Odometry (VIO) systems like VINS-Mono [10] can recover metric scale through IMU (Inertial Measurement Unit) integration, they require sufficient motion excitation and can be sensitive to initialization. Our approach provides an alternative and more direct method for metric scale recovery by leveraging LiDAR.

B. LiDAR Odometry and Geometric Degeneracy

In contrast to vision, LiDAR provides direct 3D metric measurements, eliminating scale ambiguity. LiDAR Odometry (LO) methods typically operate by registering consecutive point clouds. The seminal LOAM [5] achieved real-time performance by extracting and matching planar and edge features, a concept that inspired a generation of subsequent methods. Modern LiDAR-Inertial Odometry (LIO) systems, such as LIO-SAM [11] and FAST-LIO2 [12], achieve remarkable accuracy by tightly fusing LiDAR and IMU data.

Despite their metric accuracy, LiDAR-based methods are susceptible to failure in geometrically degenerate or sparse environments. In scenes lacking distinct structural features, such as long hallways or open fields, the registration problem becomes ill-constrained, leading to significant drift [5]. This reliance on geometric structure is a key motivation for our work, where we propose to use visual features as the primary driver for data association, side-stepping the need for robust geometric features in the environment.

C. LiDAR-Camera Fusion Strategies

To combine the complementary strengths of these sensors, various fusion strategies have been explored. Loosely-coupled methods fuse the state estimates from independent VO and LO pipelines [13], but this is sub-optimal as it does not share raw sensor information. Tightly-coupled methods are more powerful, jointly optimizing residuals from both sensors within a single estimation framework.

Early tightly-coupled systems like V-LOAM [14] demonstrated the benefits of joint optimization by using LiDAR to provide scale to a visual odometry front-end. More recent work has also focused on using LiDAR to enhance visual systems, for example, by initializing 3D landmarks for visual SLAM [15]. These approaches, however, often result in complex optimization problems or still rely on traditional visual SLAM back-ends.

Our work, **Pose360**, proposes a more focused and elegant fusion strategy. Instead of building a complex joint optimization problem, we use LiDAR for a single, critical purpose: to lift robust, pre-established 2D visual correspondences into metric 3D space. This approach is distinct from prior work in two key ways. First, by leveraging a powerful learned matcher (LightGlue) as our front-end, we decouple the correspondence-finding problem from the geometric estimation problem. Second, by transforming the problem into a simple 3D-3D alignment task, we use a direct closed-form solver, bypassing the need for iterative, non-linear optimization required by many VIO or SLAM back-ends. This makes our method simple, computationally efficient, and directly targets the core problem of metric pose recovery.

III. Methodology

The core of our **Pose360** system is a multi-stage pipeline designed to estimate the relative rigid-body transformation, $\mathbf{T}_{k \leftarrow k-1}$, between two consecutive sensor frames captured at times $k-1$ and k. This transformation consists of a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$. As illustrated in Figure III, our pipeline is organized into four main stages: (1) Data Acquisition and Pre-processing, (2) Visual Feature Extraction and Matching, (3) LiDAR-Visual Fusion for 3D Point Generation, and (4) Closed-Form Pose Estimation.

A. Data Acquisition and Pre-processing

Our system is designed for a sensor suite comprising a 360° panoramic camera and a 3D LiDAR, rigidly mounted on a mobile platform. We assume the extrinsic transformation, $\mathbf{T}_{cam \leftarrow lidar}$, which defines the rotation and translation from the LiDAR frame to the camera frame, is known and has been pre-calibrated. The sensors operate at the same frequency, and hardware-level synchronization ensures that each panoramic image I_k has a corresponding LiDAR point cloud $\mathcal{P}_k$ captured at a minimally offset timestamp.

B. Visual Feature Extraction and Matching

To establish robust visual correspondences, we process pairs of consecutive panoramic images, I_{k-1} and I_k. We employ SuperPoint [9], a deep convolutional neural network, to detect a set of distinctive 2D keypoints and their associated high-dimensional descriptors in each image. Let $\mathcal{K}_{k-1} = \{(\mathbf{u}_i, \mathbf{d}_i)\}_{i=1}^{N_{k-1}}$ and $\mathcal{K}_k = \{(\mathbf{u}_j, \mathbf{d}_j)\}_{j=1}^{N_k}$ be the sets of keypoints and descriptors for images I_{k-1} and I_k, respectively, where $\mathbf{u} = (u, v)$ represents the pixel coordinates.

These two sets of local features are then fed into a learned matcher, such as LightGlue [16] or SuperGlue [6]. These matchers leverage graph neural networks to reason about the geometric context of the entire scene, making them highly effective at rejecting outliers and matching under challenging viewpoint changes. The output is a set of high-confidence 2D-2D matches, $\mathcal{M} = \{(\mathbf{u}_i, \mathbf{u}_j) | \mathbf{u}_i \in I_{k-1}$ corresponds to $\mathbf{u}_j \in I_k\}$.

C. LiDAR-Visual Fusion for 3D Point Generation

This stage is the core of our fusion strategy, where we leverage the LiDAR data to lift the 2D visual matches into metric 3D space. For each LiDAR point cloud $\mathcal{P}_k$, we first

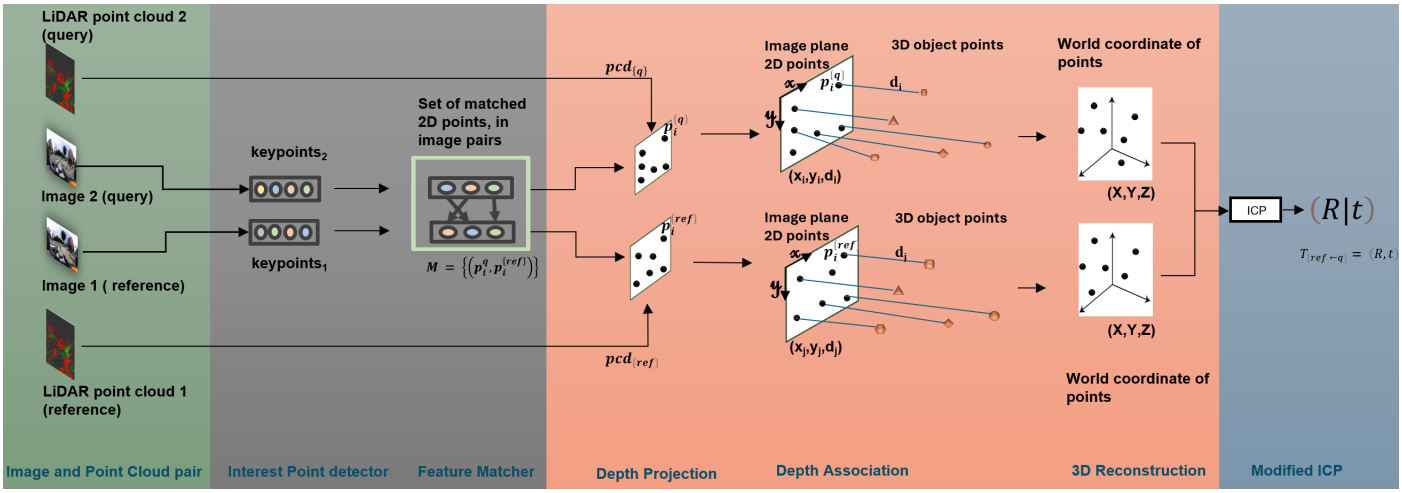

Fig. 1. T
he Lidar360Pose Pipeline. Robust 2D visual matches are established between images and lifted to sparse 3D-3D correspondences using depth from a synchronized LiDAR point cloud. An efficient, closed-form SVD-based solver then computes the final 6-DoF transformation by aligning these point sets.

create an efficient data structure for fast depth lookups. Each 3D LiDAR point $\mathbf{X}_{lidar} \in \mathcal{P}_k$ is projected into the panoramic image's coordinate system. This is achieved by first transforming the point into the camera frame using the known extrinsic calibration, $\mathbf{X}_{cam} = \mathbf{T}_{cam \leftarrow lidar}\mathbf{X}_{lidar}$, and then applying the camera's specific projection model. For the equirectangular projection used in our work, a 3D point $\mathbf{X}_{cam} = (X_c, Y_c, Z_c)$ is mapped to pixel coordinates (u, v). This process creates a sparse depth map where the depth $d = \|\mathbf{X}_{cam}\|$ is associated with the resulting pixel location (u, v).

For each 2D match $(\mathbf{u}_i, \mathbf{u}_j) \in \mathcal{M}$, we query the corresponding depth maps from frames $k-1$ and k to find the depth values d_i and d_j. We perform bilinear interpolation to find the depth at the sub-pixel keypoint location. A match is considered valid only if depth information is available for *both* keypoints, filtering out matches that fall in areas without LiDAR coverage. On average, this filtering step discards approximately 15% of the initial 2D matches, retaining a sparse but high-quality set for pose estimation. For each valid match, we back-project the 2D pixel coordinates $\mathbf{u}$ with its associated depth d to a 3D point $\mathbf{P}$ in the camera's local frame. This is achieved using the inverse of the camera's projection model. This process results in two sets of 3D points, $\{\mathbf{P}_i^{k-1}\}$ and $\{\mathbf{P}_i^k\}$, which represent the same set of physical scene points as viewed from two different camera poses.

D. Closed-Form Pose Estimation

With the two sets of 3D-3D correspondences, $\{(\mathbf{P}_i^{k-1}, \mathbf{P}_i^k)\}_{i=1}^{|\mathcal{M}|}$, the final task is to compute the relative transformation $\mathbf{T}_{k \leftarrow k-1}$ that best aligns them. This is a classic absolute orientation problem, which can be solved efficiently and non-iteratively. We seek to find the rotation $\mathbf{R}$ and translation $\mathbf{t}$ that minimize the sum of squared Euclidean distances:

$$\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{|\mathcal{M}|} \|\mathbf{P}_i^k - (\mathbf{R}\mathbf{P}_i^{k-1} + \mathbf{t})\|^2 \tag{1}$$

We employ the closed-form method of Arun et al. [17], which utilizes Singular Value Decomposition (SVD). First, we compute the centroids of both point sets, $\bar{\mathbf{P}}^{k-1}$ and $\bar{\mathbf{P}}^k$. Then, we construct the covariance matrix $\mathbf{H}$ from the centered points:

$$\mathbf{H} = \sum_{i=1}^{|\mathcal{M}|} (\mathbf{P}_i^{k-1} - \bar{\mathbf{P}}^{k-1})(\mathbf{P}_i^k - \bar{\mathbf{P}}^k)^T \tag{2}$$

We compute the SVD of $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. The optimal rotation and translation are then given by:

$$\mathbf{R} = \mathbf{V}\mathbf{U}^T \tag{3}$$

$$\mathbf{t} = \bar{\mathbf{P}}^k - \mathbf{R}\bar{\mathbf{P}}^{k-1} \tag{4}$$

A final check on the determinant of $\mathbf{R}$ is performed to handle potential reflection cases. This closed-form solution is computationally efficient and provides a highly accurate estimate of the 6-DoF relative pose, which forms the output of our odometry system.

IV. Experimental Results

We conducted a comprehensive experiment to rigorously evaluate the end-to-end performance of our **Pose360** odometry pipeline. The system was integrated onto a mobile robotic platform and tested on a long-term trajectory in a challenging, real-world environment.

A. Experimental Setup

The experiments were performed using data from a mobile robot equipped with an Ouster OS1-64 LiDAR and a Ricoh Theta Z1 panoramic camera. Our evaluation dataset was recorded at the ETRI campus in Daejeon, Republic of Korea, covering a $450\,\mathrm{m}$ trajectory through mixed indoor and outdoor environments. As a ground truth reference, we utilized a high-fidelity LiDAR SLAM system run in localization mode against a pre-built map. We evaluate the raw output of our system to assess its real-world performance.

B. Quantitative and Qualitative Analysis

To ensure a fair and accurate comparison, the estimated and ground truth trajectories were first synchronized by associating poses with the closest timestamps. An SE(3) transformation was then used to align the starting poses. All subsequent metrics are computed on these processed trajectories.

We first evaluate the global consistency of the trajectory using the Absolute Trajectory Error (ATE), which measures the direct difference between the ground truth and the estimated trajectory after alignment. The results, summarized in Table I, show an exceptionally low Root Mean Square Error (RMSE) of $0.177\,\mathrm{m}$ over the entire $450\,\mathrm{m}$ course. This quantitatively demonstrates the system's high global accuracy and minimal accumulated drift.

TABLE I
ABSOLUTE TRAJECTORY ERROR (ATE) FOR TRANSLATION

Metric	RMSE (m)	Mean (m)	Median (m)	Max (m)
ATE	0.177	0.149	0.142	0.354

To analyze the local tracking accuracy, we compute the Relative Pose Error (RPE), shown in Table II. The low mean error of approximately $8.9\,\mathrm{cm}$ confirms the system's high precision and low rate of drift.

TABLE II
RELATIVE POSE ERROR (RPE) FOR TRANSLATION (PER FRAME)

Metric	RMSE (m)	Mean (m)	Std (m)	Max (m)
RPE	0.138	0.089	0.106	0.814

Figure 2 provides a qualitative view of the system's performance. The aligned 3D trajectory in Fig. 2(a) visually confirms the high accuracy reported by the ATE, showing a remarkable structural similarity between the estimated path and the ground truth. Figure 2(b) shows the performance per-axis over time. The RPE plot in Fig. 2(c) it reveals that a initial error spike, due to a lidar slam initial localization in the start of the system, then remains consistently low, validating the effectiveness and stability of our metric-grounded visual odometry approach.

C. System Integration Validation

To demonstrate real-world applicability, we integrated **Pose360** as the odometry source within the Teach and Replay navigation framework on our mobile robots. We successfully executed a full autonomous navigation mission, validating our system's robustness and suitability for integration into complex robotic systems.

V. DISCUSSION

The experimental results provide compelling, quantitative evidence for the efficacy of our proposed fusion strategy. The low ATE RMSE of $0.177\,\mathrm{m}$ over a long-term, $450\,\mathrm{m}$ trajectory (Table I) indicates of the system's high global accuracy and minimal drift. The initial large error seen in the RPE plot is clearly shown to be an isolated artifact and does not affect the system's long-term stability.

The success of **Pose360** can be attributed to its fundamental design: it leverages a state-of-the-art learned visual matcher for robust data association and uses LiDAR for the one task it excels at—providing precise, unambiguous metric scale. This decouples the problem of finding correspondences from the problem of geometric registration. By grounding a sparse set of high-quality visual features in 3D, our system avoids the fragility of pure LiDAR odometry in geometrically sparse areas while completely resolving the scale ambiguity of pure monocular odometry. The successful integration into our navigation framework further substantiates that this approach is not just accurate, but also robust and reliable enough for use in a complete robotic application.

A. Limitations and Future Work

It is important to distinguish Pose360 as a pure odometry system, not a full Simultaneous Localization and Mapping (SLAM) system. As an odometry pipeline, it estimates query to a database or frame-to-frame motion but does not perform loop closure detection or global optimization to correct for accumulated drift over very long trajectories. While our results demonstrate low drift, this architectural distinction informs our future work.

Furthermore, despite the strong performance, our system has other limitations. Our reliance on a visual front-end means that performance will inevitably degrade in globally textureless environments or under extreme weather conditions where the feature matcher fails. In scenarios with dense fog or heavy rain, where both the camera and LiDAR data may be compromised, the system's accuracy would be significantly impacted. While the use of learned features provides a degree of robustness to moderate lighting changes, severe visual degradation remains a challenge. This motivates our primary direction for future research: extending the odometry pipeline to a full SLAM system. This would involve:

1) **Pose Graph Optimization:** Developing a back-end that incorporates our V-LIO pose estimates as factors in a pose graph. This would allow for the integration of other odometry sources (e.g., IMU, wheel encoders) as additional factors for improved robustness.
2) **Loop Closure Detection:** Integrating a place recognition module to detect when the robot has returned to a previously visited area. Adding these loop closure constraints to the pose graph would allow the system

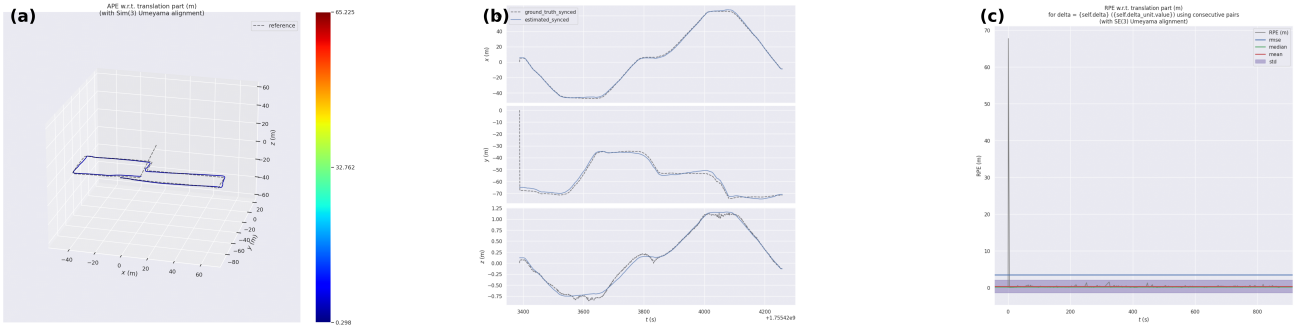

Fig. 2. Trajectory evaluation results. **(a)** The aligned 3D trajectory shows high global accuracy against the ground truth. **(b)** Per-axis plots confirm close tracking over time. **(c)** The RPE plot reveals an initial artifact, followed by consistently low local error, demonstrating the system's stability.

to correct for any accumulated drift, enabling true long-term SLAM.

3) **Dynamic Environment Handling:** Incorporating robust estimation techniques or a dynamic object segmentation front-end to identify and reject features that lie on moving objects, a known challenge for all visual odometry systems.

VI. Conclusion

In this paper, we presented **Lidar360Pose**, a novel Visual-LiDAR odometry system that resolves the inherent scale ambiguity of visual methods by grounding 2D learned feature matches with precise 3D depth from a synchronized LiDAR. Our approach transforms the pose estimation problem into an efficient, closed-form 3D alignment task. We demonstrated through a rigorous real-world experiment on a mobile robot that our system achieves high accuracy and global consistency, attaining a translation ATE of just $0.177\,\mathrm{m}$ RMSE when evaluated against a high-fidelity LiDAR SLAM ground truth. Furthermore, the successful integration of our system into a full-scale robotic navigation stack validates its practical utility and robustness. By elegantly fusing the complementary strengths of vision and LiDAR, Lidar360Pose offers an effective and reliable solution for metric state estimation, a critical enabler for autonomous systems in real-world environments.

VII. Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00215760, Guide Dog: Development of Navigation AI Technology of a Guidance Robot for the Visually Impaired Person).

References

[1] C. Cadena et al., 'Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age', IEEE Transactions on robotics, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] S. Thrun, W. Burgard, and D. Fox, Probabilistic robotics. MIT press, 2005.

[3] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, 'ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial and Multi-Map SLAM', in IEEE Transactions on Robotics, 2021, vol. 37, pp. 1874–1890.

[4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, 'MonoSLAM: Real-time single camera SLAM', IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 1052–1067, 2007.

[5] J. Zhang and S. Singh, 'LOAM: Lidar Odometry and Mapping in Real-time', Robotics: Science and Systems, vol. 2, no. 1, pp. 1–9, 2014.

[6] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, 'SuperGlue: Learning Feature Matching with Graph Neural Networks', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4938–4947.

[7] D. G. Lowe, 'Distinctive image features from scale-invariant keypoints', International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, 'ORB: An efficient alternative to SIFT or SURF', in 2011 International conference on computer vision, 2011, pp. 2564–2571.

[9] D. DeTone, T. Malisiewicz, and A. Rabinovich, 'SuperPoint: Self-Supervised Interest Point Detection and Description', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 224–236.

[10] T. Qin, P. Li, and S. Shen, 'VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator', in IEEE Transactions on Robotics, 2018, vol. 34, pp. 1004–1020.

[11] T. Shan, B. Englot, C. Ratti, and D. Rus, 'LIO-SAM: Tightly-coupled Lidar Inertial Odometry via Smoothing and Mapping', in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 5135–5142.

[12] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, 'Fast-lio2: Fast direct lidar-inertial odometry', IEEE Transactions on Robotics, vol. 38, no. 4, pp. 2053–2073, 2022.

[13] J. Zhang, W. Wen, F. Huang, X. Chen, and L.-T. Hsu, 'Coarse-to-fine loosely-coupled lidar-inertial odometry for urban positioning and mapping', Remote Sensing, vol. 13, no. 12, p. 2371, 2021.

[14] J. Zhang and S. Singh, 'Visual-lidar odometry and mapping: Low-drift, robust, and fast', in 2015 IEEE international conference on robotics and automation (ICRA), 2015, pp. 2174–2181.

[15] S. Wang, Y. Kobayashi, A. A. Ravankar, A. Ravankar, and T. Emaru, 'A novel approach for lidar-based robot localization in a scale-drifted map constructed using monocular slam', Sensors, vol. 19, no. 10, p. 2230, 2019.

[16] P. Lindenberger, P.-E. Sarlin, and S. D'Aronco, 'LightGlue: Local Feature Matching at Light Speed', in International Conference on Computer Vision (ICCV), 2023.

[17] K. S. Arun, T. S. Huang, and S. D. Blostein, 'Least-squares fitting of two 3-D point sets', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-9, no. 5, pp. 698–700, 1987.