From 2D to 3D: AISG-SLA Visual Localization Challenge

Jialin Gao[1] , **Bill Ong**[1] , **Darld Lwi**[2] , **Zhen Hao Ng**[2] , **Xun Wei Yee**[1] , **Mun-Thye Mak**[1] , **Wee Siong Ng**[3] , **See-Kiong Ng**[4] , **Hui Ying Teo**[2] , **Victor Khoo**[2] , **Georg Bökman**[5] , **Johan Edstedt**[6] , **Kirill Brodt**[7] , **Clémentin Boittiaux**[8] , **Maxime Ferrera**[8] and **Stepan Konev**[9]

[1]AI Singapore, Singapore
[2]Singapore Land Authority, Singapore
[3]Institute for Infocomm Research, Singapore
[4]National Univerity of Singapore, Singapore
[5]Chalmers University of Technology, Sweden
[6]Linköping University, Sweden
[7]Université de Montréal, Canada
[8]Ifremer, Centre Méditerranée, France
[9]Booking.com, Netherlands
{jialin, bill_ong}@aisingapore.org, {darld_lwi, ng_zhen_hao}@sla.gov.sg, {xunwei, munthye}@aisingapore.org, wsng@i2r.a-star.edu.sg, seekiong@nus.edu.sg, {teo_hui_ying, vic-tor_khoo}@sla.gov.sg, bokman@chalmers.se, johan.edstedt@liu.se, kirill.brodt@umontreal.ca, {boittiauxclementin, maxime.ferrera, stevenkonev}@gmail.com

Abstract

Research in 3D mapping is crucial for smart city applications, yet the cost of acquiring 3D data often hinders progress. Visual localization, particularly monocular camera position estimation, offers a solution by determining the camera's pose solely through visual cues. However, this task is challenging due to limited data from a single camera. To tackle these challenges, we organized the AISG–SLA Visual Localization Challenge (VLC) at IJCAI 2023 to explore how AI can accurately extract camera pose data from 2D images in 3D space. The challenge attracted over 300 participants worldwide, forming 50+ teams. Winning teams achieved high accuracy in pose estimation using images from a car-mounted camera with low frame rates. The VLC dataset is available for research purposes upon request via *vlc-dataset@aisingapore.org*.

1 Introduction

Camera pose estimation, also referred to as visual localization [Barros *et al.*, 2022], plays a pivotal role in determining the 6-degree-of-freedom pose (3D position and orientation) of a camera within its environment using visual cues. Specifically, monocular camera poses estimation [Engel *et al.*, 2014] involves utilizing a single camera to derive the camera's position and orientation. This process entails analyzing 2D images captured by the camera and computing a transformation matrix that characterizes the camera's spatial relationship to the 3D world. Despite the challenges posed by the limited information available from a single camera, the significance of camera pose estimation cannot be overstated, especially considering the high costs associated with obtaining 3D data from alternative sources like LiDAR sensors.

Accurate camera pose estimation is indispensable across a spectrum of real-world applications, ranging from urban planning [Coors *et al.*, 2000] to augmented reality, underwater surveillance [González-Sabbagh and Robles-Kelly, 2023], robotics, and autonomous vehicles. In urban planning, camera pose estimation equips planners with valuable spatial insights that inform decision-making processes. In augmented reality, precise knowledge of the user's device position and orientation relative to the surrounding environment is pivotal for convincingly overlaying virtual objects. Moreover, camera pose estimation plays a critical role in underwater surveillance by facilitating the tracking and monitoring of objects, structures, and environmental conditions in aquatic environments. Furthermore, in robotics and autonomous vehicles, accurate camera pose estimation empowers robots to navigate and interact with their surroundings effectively.

To promote the development and testing of models in more dynamic and diverse environments, we introduce the AISG–SLA Visual Localization Challenge (VLC). The primary objective of the challenge is to discern the relative pose estimates among images obtained by a monocular camera affixed to a vehicle navigating the streets of Singapore. The focus lies in accurately measuring the rotational differences between successive images, reflecting the subtle variations in orientation as the vehicle progresses along its path. In addition, the secondary objective involves precisely estimating relative translations, and capturing the spatial movements and displacements between scenes observed in consecutive frames. These two assessment criteria require participants to address the complexities of visual localization in dynamic urban settings effectively for both rotational alignment and positional tracking within real-world driving scenarios.

The paper's structure is as follows: Section 2 reviews existing camera pose estimation studies. Section 3 explores the VLC dataset's characteristics and challenges. Section 4 outlines winning teams' strategies [1]. The conclusion offers final remarks.

2 Related Work

Within the realm of camera pose estimation, there exist two primary approaches for solving the associated problems: direct [Engel *et al.*, 2017] and indirect [Mur-Artal *et al.*, 2015] methods. Indirect methods typically involve detecting interest points, associating them with feature descriptors, and optimizing the camera pose and 3D point clouds by minimizing reprojection error [Rosinol *et al.*, 2020; Campos *et al.*, 2021]. On the other hand, direct methods delve into the image formation process and define objectives based on photometric error [Zubizarreta *et al.*, 2020]. While direct methods capture more image details, such as lines and intensity variations [Engel *et al.*, 2017], they face more complex optimization challenges and are less robust to geometric distortions.

Recent advancements in deep learning-based techniques [Zhou *et al.*, 2018] have shown promise in addressing challenging scenarios in camera pose estimation. These techniques often involve training systems for specific subtasks, such as feature detection, matching, and localization. However, some deep learning models tend to concentrate on small-scale reconstruction and may lack capabilities like loop closure and global bundle adjustment, limiting their applicability for large-scale deployment.

Current methodologies heavily rely on training deep models using datasets such as KITTI [Geiger *et al.*, 2013], EuRoC [Burri *et al.*, 2016], and TartanAir [Wang *et al.*, 2020]. However, these datasets often exhibit limited motion amplitude and may contain synthetic data, which may not accurately represent real-world application scenarios.

3 VLC Dataset

In this section, we outline the characteristics of the dataset.

3.1 Data Sources

The dataset utilized in this research comprises photographs taken by a spherical camera system equipped with a 5-megapixel resolution and a Sony IMX264 $2/3''$ CMOS sensor, featuring a pixel size of $3.45 \times 10 - 3$ mm and utilizing 4.4 mm focal length lenses. These photographs were captured within the urban areas of two different townships in Singapore, primarily showcasing human subjects (faces captured in images were blurred), vehicles, architectural structures, and natural landscapes predominantly found within residential areas and parks.

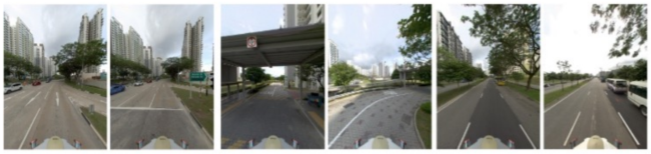

Figure 1: The camera is directed towards the rear of the vehicle, resulting in consecutive captures where the movement is not depicted within the image frame.

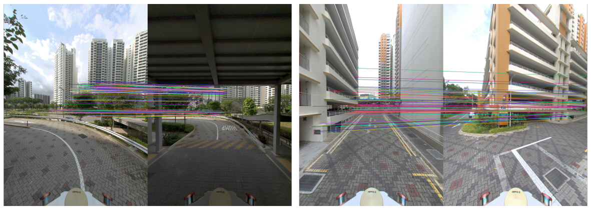

Figure 2: Illustration of challenging matches by LightGlu.

The images were captured based on distance (i.e., every 10 meters), estimated by the number of tire rotations. This is approximately equivalent to 1 or 2 shots every second. Some distances between timestamps might have a larger difference (e.g., 8.8 meters vs. 10 meters) because the car might have been making a turn. The visual data for the VLC was provided in the form of street-level monocular images stored in JPEG format. The dataset was divided into two sets: a training set containing 10,007 images and a testing set containing 2,219 images. These sets were further subdivided into trajectories, with four trajectories allocated for training purposes and one trajectory reserved for testing. Here, a trajectory refers to a sequence of images captured as the camera moves along a continuous path. The intrinsic parameters, including the focal length and optical center (also known as the principal point), were provided.

3.2 Data Matching Challenge

Several challenges were introduced into the dataset to enhance its complexity. First, there were considerable variations in the time intervals between successive frames, with some intervals spanning up to 1.5 minutes, resulting in limited overlapping visual features. Secondly, the frame rate differed significantly across different trajectories, leading to noticeable discontinuities in the dataset. To streamline the dataset, the lower portion of each image, which predominantly depicted static car components, was excluded from consideration as it was deemed irrelevant for pose estimation tasks. Furthermore, the dataset exhibited diverse lighting conditions, necessitating that models developed by participants demonstrate robustness in handling such variations. Figure 1 illustrates the challenges within the dataset. Firstly, on the left, we have a standard pair of images (1 sec apart) that are relatively easy to match. In the center is a typical pair (2 sec apart) that presents more difficulty in matching due to sharp turns, which occur infrequently in the sequence, making them particularly challenging. Finally, on the right is a pair of images (76 sec apart) that can hardly be matched.

Despite being consecutive in the sequence, some images

[1] https://prizechallenge.aisingapore.org/competitions/1/visual-localisation/leaderboard/

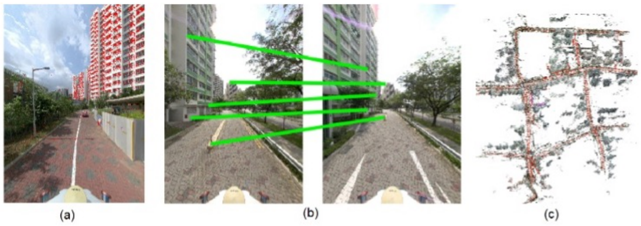

Figure 3: Method pipeline. 3(a) Extract DeDoDe keypoints in all images and fine non-sequential image pairs to match using i) DINO v2, ii) manual inspection. 3(b) Match keypoints in sequential and non-sequential image pairs using RoMa, filter with Graph-Cut RANSAC. 3(c) Structure from motion using COLMAP, each red dot is a position where a picture was taken.

display characteristics such as time gaps in the dataset, where non-overlapping pairs may result in the method generating arbitrary relative poses, leading to subpar outcomes in straightforward sequential matching. In Figure 2, we illustrate challenging matches addressed by LightGlue [Lindenberger *et al.*, 2023], a deep neural network tailored to matching sparse local features between image pairs. In the left image, numerous matching points are obscured by the roof, while in the right image, a 90° turn of the car causes matching failure.

4 Proposed Methods

The AISG–SLA Visual Localization Challenge took place at IJCAI 2023 from May 26, 2023, to July 26, 2023, featuring a total prize pool of up to USD 40,000. With over 300 participants globally, the event saw the formation of more than 50 teams. In this section, we provide an overview of the winning teams' strategies and their approaches to addressing the pose estimation problem. Due to space limitations, readers are encouraged to contact the teams for further details [2].

4.1 RoMa and DeDoDe Strategy

GETINGARNA, the first-place winning team, scored 0.0273 in rotational error and 1.4205 in translational error, leverages their recent research on deep learning-based image matching, notably the RoMa [Edstedt *et al.*, 2023] and DeDoDe methods [Edstedt *et al.*, 2024]. These techniques excel in reliably estimating relative poses for most consecutive image pairs within the challenge sequence.

To attain accurate estimates for the most challenging image pairs and address loop closures. The team devised a pipeline that integrates image retrieval with DINOv2 [Oquab *et al.*, 2023], and full structure-from-motion reconstruction with COLMAP [Schönberger and Frahm, 2016; Schönberger *et al.*, 2016]. By employing image retrieval with DINOv2, they successfully matched specific non-consecutive image pairs, effectively addressing the challenge of significant time gaps in the image sequence. These improvements contributed to their solution outperforming all competitors.

KBRODT, securing 2nd place (Figure 4), adopted a similar approach, leveraging RoMa and the deep neural network model DeDoDe descriptor upon recognizing the limitations of ORB detectors [Campos *et al.*, 2021] and FLANN matchers [Muja and Lowe, 2009] for the task at hand. Their strategy revolves around utilizing publicly available pretrained weights for both RoMa and DeDoDe descriptor models. This strategic decision not only streamlined their implementation process but also eliminated the necessity for extensive retraining, ensuring stable and robust performance across various tasks. Notably, the team achieved a rotational error of 0.0382 and a translational error of 6.3374.

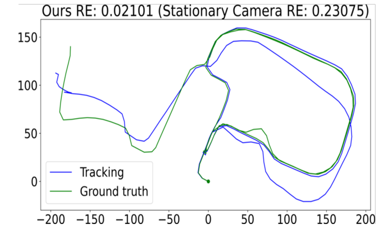

Figure 4: Visual representation of predicted sub-trajectories of the rotational error (RE).

4.2 CNN-based Strategy

The third-place winning teams, Team SYLISH and Team SDRNR (tied), both adopted a CNN-based strategy in their architectural designs. SYLISH's solution integrated LightGlue and MobileNetV2 [Sandler *et al.*, 2018] as fundamental components, establishing a deep image retrieval network and deep feature matching for 2D-2D correspondence among similar images. They inferred relative motion between frames through essential matrix estimation within a RANSAC scheme and estimated translation scale factors using a deep-scale estimation network. The resulting poses underwent refinement via nonlinear least-squares optimization to address rotation graph problems.

SDRNR utilized basic CNN approaches like EfficientNet and ResNet. These models processed two RGB images from consecutive states, accompanied by precomputed depth maps and flow maps, predicting quaternions for rotation and corresponding translations. They advanced their model by integrating a visual matching system capable of identifying anchor point matches on RGB images, even amidst significant camera pose changes. Additionally, they implemented heuristics to address situations where no matches were found, particularly during 180° turns or anomalously long time deltas between states.

5 Conclusion

3D mapping research is crucial for smart city development, but obtaining 3D data is expensive. Monocular camera position estimation provides a cost-effective solution by determining camera pose from visual cues. The AISG–SLA Visual Localization Challenge (VLC) at IJCAI 2023 showed that state-of-the-art techniques from the research community can be effective in accurately extracting camera pose data from 2D images in the real world. Furthermore, the VLC dataset, available for research purposes, provides a valuable resource for further research in dynamic urban environments.

[2] https://aisingapore.org/aisg-sla-visual-localisation-challenge-winners/

Acknowledgements

AISG - This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme.

Team Getingarna (Georg Bökman, Johan Edstedt) - This work was supported by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation and by the strategic research environment ELLIIT funded by the Swedish government.

Team kbrodt (Kirill Brodt) - We acknowledge Novosibirsk State University for providing the GPU server.

References

[Barros *et al.*, 2022] Andréa Macario Barros, Maugan Michel, Yoann Moline, Gwenolé Corre, and Frédérick Carrel. A comprehensive survey of visual slam algorithms. *Robotics*, 11:24, 2022.

[Burri *et al.*, 2016] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

[Campos *et al.*, 2021] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[Coors *et al.*, 2000] Volker Coors, Tassilo Huch, and Ursula Kretschmer. Matching buildings: Pose estimation in an urban environment. In *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*, pages 89–92. IEEE, 2000.

[Edstedt *et al.*, 2023] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Revisiting robust losses for dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023.

[Edstedt *et al.*, 2024] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024.

[Engel *et al.*, 2014] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[Engel *et al.*, 2017] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[González-Sabbagh and Robles-Kelly, 2023] Salma P González-Sabbagh and Antonio Robles-Kelly. A survey on underwater computer vision. *ACM Computing Surveys*, 55(13s):1–39, 2023.

[Lindenberger *et al.*, 2023] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023.

[Muja and Lowe, 2009] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Rosinol *et al.*, 2020] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696. IEEE, 2020.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[Schönberger and Frahm, 2016] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 4104–4113. IEEE, 2016.

[Schönberger *et al.*, 2016] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016.

[Wang *et al.*, 2020] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[Zhou *et al.*, 2018] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 822–838, 2018.

[Zubizarreta *et al.*, 2020] Jon Zubizarreta, Iker Aguinaga, and Jose Maria Martinez Montiel. Direct sparse mapping. *IEEE Transactions on Robotics*, 36(4):1363–1370, 2020.