
Proceedings Paper:
Sun, L. orcid.org/0000-0002-0393-8665, Taher, M., Wild, C. et al. (6 more authors) (2021) Robust and long-term monocular teach-and-repeat navigation using a single-experience map. In: Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 27 Sep - 01 Oct 2021, Virtual conference (Prague, Czech Republic). IEEE , pp. 2635-2642. ISBN 9781665417150

https://doi.org/10.1109/IROS51168.2021.9635886

Robust and Long-term Monocular Teach and Repeat Navigation using a Single-experience Map

Li Sun[1*], Marwan Taher[1], Christopher Wild[1], Cheng Zhao[2], Filip Majer[3], Zhi Yan[4], Tomáš Krajník[3], Tony Prescott[1] and Tom Duckett[5]

***Abstract*— This paper presents a robust monocular visual teach-and-repeat (VT&R) navigation system for long-term operation in outdoor environments. The approach leverages deep-learned descriptors to deal with the high illumination variance of the real world. In particular, a tailored self-supervised descriptor, DarkPoint, is proposed for autonomous navigation in outdoor environments. We seamlessly integrate the localisation with control, in which proportional–integral control is used to eliminate the visual error with the pitfall of the unknown depth. Consequently, our approach achieves day-to-night navigation using a single-experience map and is able to repeat complex and fast manoeuvres. To verify our approach, we performed a vast array of navigation experiments in various outdoor environments, where both navigation accuracy and robustness of the proposed system are investigated. The experimental results show that our approach is superior to the baseline method with regards to accuracy and robustness.**

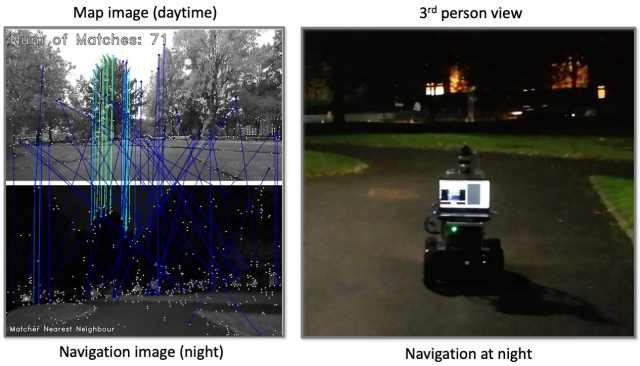


Fig. 1: Our navigation approach is able to operate from day to night using a map created in a single teaching session.

I. INTRODUCTION

Vision-based navigation has the potential to be mass produced utilising the benefits of low-cost cameras, cheap computation, and novel machine learning paradigms. However, unlike the commonly used 2D and 3D lidars, cameras are naturally passive sensors. Thus, camera-based mapping and localisation systems are prone to illumination changes, feature deficiency situations, and appearance variations. The resulting reliability issues mean that using vision to create detailed, globally consistent maps of large areas can be a very difficult task. However, several vision-based teach-and-repeat navigation systems do not rely on global map consistency, which are capable of reliably following previously-taught trajectories organised in a topological map [1], [2], [3].

The main challenges in perception and localisation for vision-based navigation are two-fold. Firstly, the visual features need to be robust to deal with the illumination changes from day to night in order to operate in the long term. Secondly, the localisation and control should be seamlessly integrated to eliminate latencies in decision making. Most of the existing long-term vision-based navigation systems build a multi-experience map [4], [5], [6] or learn robust descriptors or representations from multi-experience navigation [7], [8], [9], [10], [11]. Using a monocular sensor and a single-experience map to enable the robot to navigate autonomously from day to night is an open research problem, which is highly demanded by outdoor robotic applications such as logistics and transportation (e.g. last-mile delivery), and environmental exploration (e.g. sample return).

To tackle these challenges, we propose to use a tailored deep local descriptor to deal with the illumination variance for long-term navigation. We also propose a seamlessly-coupled vision-guided control mechanism to rapidly integrate the localisation error and adapt to environments of various scales. The main contributions of this paper are:

- Design and implementation of a monocular teach-and-repeat navigation system using off-the-shelf camera, enabling a wheeled robot to repeat complex and dynamic manoeuvres from day to night using a single-experience map in outdoor environments;
- Demonstration and evaluation of the integrated system with an actual robotic platform, showing real-time perception-action loop (above 25Hz), accurate path following and robust localisation in large-scale environments;
- Open source work, i.e. the developed system, evaluation toolboxes as well as video demos are all available on our project website: `https://github.com/FAIRSpace-AdMaLL`.

II. RELATED WORK

Unlike active sensor-based SLAM methods, that can directly recover the environment structure, the visual (especially monocular) SLAM is affected by the appearance changes caused by varying illumination. This results in

∗ Corresponding Author: `li.sun@sheffield.ac.uk`

This project is funded by EPSRC FAIR-SPACE Hub (EP/R026092/1), EU Horizon 2020 ILIAD (No 732737) and CSF/NRF project ToLTATempo 20-27034J.

[1] Sheffield Robotics, University of Sheffield, UK
[2] Department of Engineering Science, University of Oxford, UK
[3] Czech Technical University in Prague, Czech Republic
[4] CIAD UMR7533, Univ. Bourgogne Franche-Comté, UTBM, France
[5] L-CAS, University of Lincoln, UK

reliability issues, especially in case the robots have to operate over extended periods of time [12].

One of the usages of visual SLAM for robot navigation is teach and repeat, where the robot creates a map during a teleoperated drive. This map is then used by the robot to autonomously repeat the taught trajectory later on [13], [1], [14], [2] and [15]. The principal advantage of teach-and-repeat systems is that they do not have to rely on globally consistent 2D or 3D metric maps of the environment as teach-and-repeat does not require explicit localisation [2], [15]. Instead, the navigation task in teach-and-repeat systems can be formulated as visual servoing [16], [2]. For example, [17], [18], [19] create a visual path, which is a set of images along the human-guided route, and then employ visual servoing to guide robots across the locations these images were captured at. Similarly, [16] represents the path as consecutive nodes, each containing a set of salient visual features, and uses local feature tracking to determine the robot's steering to guide it to the next node. The authors of [20] extract salient features from the video feed on the fly and associate these with different segments of the teleoperated path. When navigating a given segment, their robot moves forward and steers left or right based on the positions of the currently recognised and already mapped features. The segment end is detected by means of comparing the mapped segment's last image with the current view. While visual teach-and-repeat methods do not rely on globally consistent metric maps, they can be combined with high-level topological maps to allow path planning [3].

To cope with appearance variations in long-term deployments, visual teach-and-repeat systems have often been extended by approaches that were previously aimed at visual localisation and place recognition in changing environments such as the multi-experience framework [4], frequency map enhancement [21], feature selection [22], [23] or feature training schemes [24]. For example, in [5], [4], [6], multi-run experiences are leveraged to build a location graph where multiple appearances are stored for the same location, and incremental mapping is implemented when the localisation confidence is low. [11] proposed an adaptive map for day-to-night operation that automatically selects effective features given the temporal context, removes obsolete features and adds new ones. In [9], [10], evolutionary methods are used to select patterns from multi-session experiences for binary features, thereby enabling their long-term deployment in cross-seasonal or changing illumination conditions.

Compared to hand-crafted visual features [25], [26], the emerging deep-learned visual features [27], [28], [29], [30] show proven effectiveness in dealing with illumination changes. One of the main challenges in learning local descriptors is to associate pixels from images captured with different illumination and seasons. In D2-Net [27], graph-based Structure-from-Motion is used to associate images of the same place and 3D-to-2D projection is used to build the pixel-wise correspondence between images. Self-supervised methods such as SuperPoint [29] leverage photometric and homographic adaptation that can generate correspondences from a single image without the need of data association. SuperGlue [31] proposed a attentional Graph Neural Network to learn local feature aggregations and perform end-to-end matching.

III. METHODOLOGY

In the visual teach-and-repeat problem, given the observation o during the teaching session, a topological map o_m and action event map E_m will be built. Specifically, as shown in Fig. 2, the topological map o_m contains a sequence of images with a fixed traversal interval $\{I_m^i\}_N$, and the event map consists of a number of velocity events $\{vel_l^d, vel_a^d, \Delta_t\}_K$ which represent applying linear and angular velocities at distance d for a duration of Δ_t. The sequence can then be repeated using dead reckoning to replay the velocity events. To make the navigation precise and scalable, in the navigation (i.e. repeat) phase, a visual offset e^d can be estimated between the online observation o_t^d and the paired map observation o_m^d, then the controller will apply the velocity compensation vel_e to minimise the visual offset.

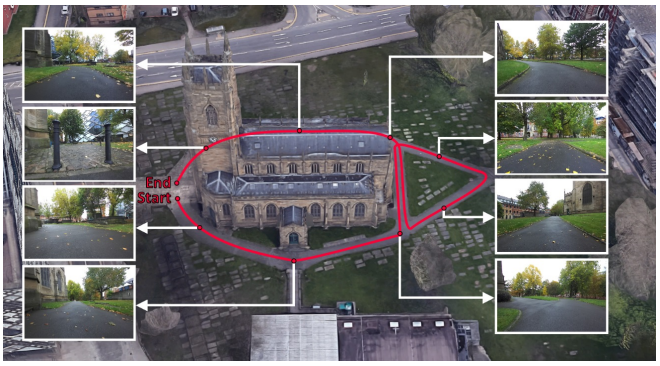


Fig. 2: An example of the topological map used in our experiment.

A. Topological Mapping

In the teach phase, explicit metric mapping and precise localisation is not required, instead, an image-based topological map is used. During this phase, the robot is driven manually via tele-operation. The wheel encoders i.e. robot odometry) can be used to estimate the distance traversed and keyframe images with a fixed translational interval[1] are saved as topological nodes. Similarly to [15], a sequence of tele-operation events is recorded as the event map (also known as "path profile"). Unlike [15], we also record the exact duration of these events to handle complex manoeuvres.

B. DarkPoint: Deep Learned Visual Descriptor

Day-to-night robot navigation requires local descriptors that are robust to dramatic illumination changes in order to register the paired day and night images. Apart from this, other desirable properties such as scale-invariance and rotation-invariance should also be achieved. The self-supervised learning of descriptors is an appropriate scheme

[1] 0.2m is used in this paper.

which leverages photometric and homographic adaptation to generate correspondences and contrastive loss for descriptor learning. In this paper, we use the VGG-like architecture similar with SuperPoint [29] due to its run-time performance[2]. Our approach, named DarkPoint, is a tailored approach for long-term teach-and-repeat navigation. We particularly strengthen illumination adaptation, in which non-linear image illumination augmentation approach, i.e. Gamma transform, is adapted to synthesise realistic images with extremely high- or low-illumination.

1) Illumination Adaption using Gamma Transform: In the complex real-world, the lightness changes non-linearly and this change is only reflected in the lightness channel of an image. The Gamma transform is a non-linear method to strengthen low-illumination images. Our intuition is to apply the Gamma transform on normal (daytime) images to generate low-illumination (night-time) images. Directly applying the illumination transform to RGB images introduces distortions in the colour channels of the correlated image. To preserve the colour information, we first transform the image to HSV (Hue, Saturation, Value) space, and a parameter γ is used to adjust the value of channel I_v (i.e. lightness) exponentially:

$$I_v^{'} = \min\left[\delta\, I_v^{\gamma^{-1}},\ \delta\right] \tag{1}$$

where δ is the margin to cap the maximum lightness values to improve the non-linearity. The adjusted channel value together with the original hue and saturation channels will be converted to a grey image for joint training. Apart from the Gamma transform, other widely used photometric augmentations, i.e. additive Gaussian noise, additive speckle noise, random contrast and additive shade, are randomly included to form the final illumination adaptation $\mathcal{P}$.

2) Joint Training: We use the base detector in SuperPoint (i.e. MagicPoint) to annotate key points on training images. In the joint training step (shown in Fig. 3), we first apply a random illumination transform $\mathcal{P}$ to the original training image to generate paired images with different illuminations. Both of them will have other photometric augmentations and one of them will be warped by applying homographic transforms. Similar to [29], homographic adaptations $\mathcal{P}$, including translation, scale, in-plane rotation and symmetric perspective distortion, are implemented for generation of correspondences. Dense pixel-to-pixel correspondences can be automatically generated using the geometric consistency:

$$f(I) = \mathcal{H}^{-1} f\big(\mathcal{H}(\mathcal{P}(I))\big) \tag{2}$$

Similar to [27], [29], the cross-entropy loss is applied on detection scores to learn the keypoint detector. While in contrast, we apply triplet loss [32] on uniformly sampled features for contrastive descriptor learning. Given a local feature triple that consists of a positive pair (d_q, d_p) and a negative pair (d_q, d_n), we minimise the metric distance between positive pairs and maximise that between negative pairs with margins:

$$\mathcal{L}_d(d_q, d_p, d_n) = [\alpha - d_q^T d_p)]_+ + [d_q^T d_n - \beta]_+ \tag{3}$$

where
$$d_q \in f(I),\ (d_p, d_n) \in \mathcal{H}^{-1} f\big(\mathcal{H}(\mathcal{P}(I))\big) \tag{4}$$

and
$$||d_q||_2 = ||d_p||_2 = ||d_n||_2 = 1. \tag{5}$$

In our implementation, the positive margin α and negative margin β are set as 1.0 and 0.2 respectively. The COCO dataset[3] is used for joint training, a γ value is randomly sampled between 0.2 and 4, which covers the variety of lightness and illumination in our application, and δ is between 0.6 and 1.0.

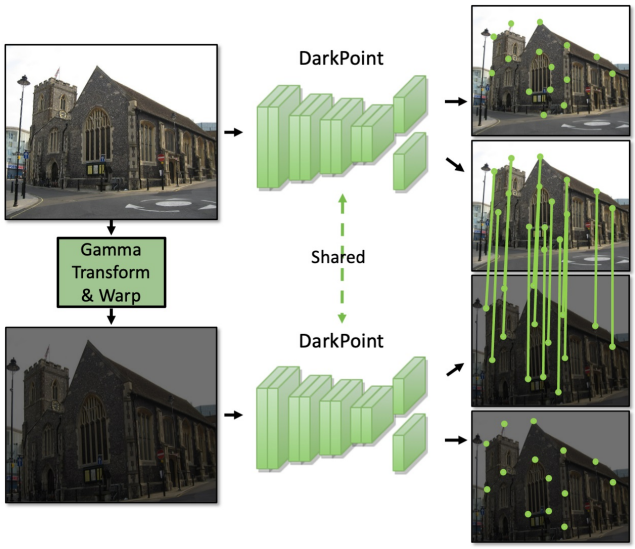


Fig. 3: In each iteration of joint training, we apply the Gamma transform with random γ and δ to syntheses high- and low-illumination images.

C. Steering Estimation and Control

In the navigation (repeat) phase, the recorded action events will be replayed, and simultaneously, at each distance d, the map image o_m^d will be loaded and matched with the corresponding online image o_t^d captured by the on-board camera to correct the relative pose. Most monocular SLAM approaches use RANSAC and scale propagation to estimate the 6DoF relative pose. However, these methods are not robust when the feature matching is extremely poor, and the scale-drifting problem cannot be mitigated.

Given a robot teach-and-repeat navigation model as described in [15], from the stereo geometry, the translational error t_y perpendicular to the teach trajectory can be approximated by:

$$t_y \propto z_t\ \mathrm{e}_{\mathrm{t}}^{\mathrm{d}}(o_m^d, o_t^d), \tag{6}$$

where $\mathrm{e}_{\mathrm{t}}(o_m, o_t)$ is the disparity (also known as visual offset) between map and camera images, z_t is the distance to the keypoint. We investigated the use of scale propagation to

[2]The on-board computer of our robot platform has a NVIDIA Geoforce RTX 2070 GPU.

[3]https://cocodataset.org/

estimate the depth of keypoint z_t from previous images, and the estimated scale drifts over time. We also tried to implement a single image depth prediction network, however, the run-time performance was not satisfactory due to the bulky decoder architecture.

In [15], histogram voting method is used to estimate the visual offset $\mathrm{e}_\mathrm{t}^\mathrm{d}$ and a convergence theorem is provided. In our approach, we also use histogram voting to estimate the visual offset without knowing the scale, and we integrate the visual offset through the time to minimise the steady state error, thereby accelerating the robot repeat convergence towards the teaching trajectory. To be more specific, a brute-force nearest neighbour matching (NN) is used for feature matching. We calculate the pixel offset in x- and y-direction for each match. Outliers with large y offset will be rejected first with the prior that the $roll$ and $pitch$ angle should be small when the robot traverses the same location. Then, the histogram voting is applied to the x offsets and the final inliers can be obtained by calculating the mean x offset value of the matches falling into the largest bin. Once visual offset e_t is estimated, the integral of the latency errors accumulates to adapt the steering control to difference scales z. The memorised angular velocity at distance d can be corrected by adding a visually-guided compensation vel_{gain}:

$$\begin{aligned} vel_a^{d'} &= vel_a^d + vel_{gain} \\ vel_{gain} &= \Phi(e_t^d + \lambda \int_0^{t-1} e_t^d \, dt), \end{aligned} \tag{7}$$

where the hyper-parameter λ is the weights of latency error integration. Φ is the weight of vision intervention. Noting that we tuned these parameters manually according to practical experience. Then we fuse the recorded teach events with the vision-guided velocity compensation for a shared control between memorised experience and vision corrections. Finally, the robot will actively localise and incrementally approximate the teaching trajectory over time.

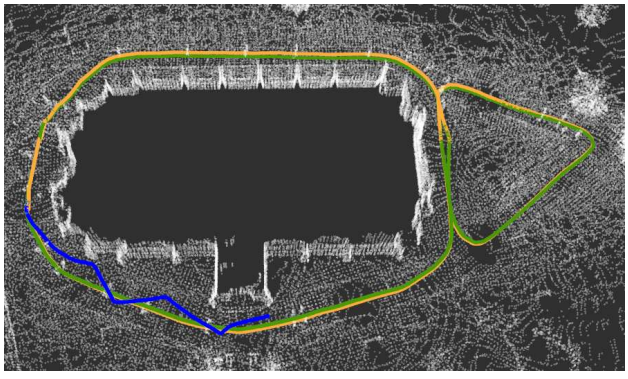


Fig. 4: An example of associated trajectories using sensor fusion. Note, other sensors, e.g. lidar, are only used for measuring the navigation error.

IV. Experiments

A. The Mobile Platform

In this paper, we use a DrRobot Jaguar 4x4 platform for the experiments. It is an all-terrain robot, with dimensions of $62cm \times 57cm \times 90cm$ and a weight of 35KG. The robot is equipped with a ZED2 stereo camera, a 3D Ouster OS1-64 lidar, a Xsens MTi-G710-GNSS and a Dell G5 laptop. Utilising a battery bank, the robot with the computational devices is able to operate for around four hours. Our navigation approach only requires a monocular camera. We use one of the ZED2 stereo cameras as a standard monocular camera in our experiments. Lidar-inertial-GPS SLAM and ICP are used for tracking and association of multi-session trajectories, to generate the "ground-truth" for the evaluation of navigation performance (see Fig. 4).

B. Evaluation Metrics

Autonomous vision-based navigation involves an active localisation process which needs to be accurate and robust. Two metrics are used to evaluate the navigation accuracy and navigation robustness:

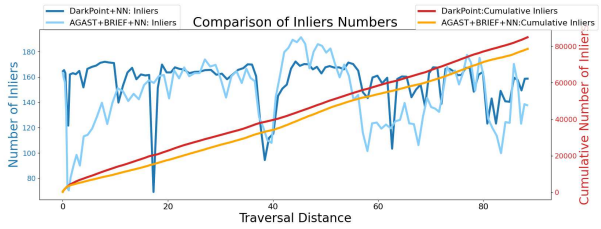
1) Navigation Accuracy: the absolute trajectory error (ATE) and the relative pose error (RPE), including RMSE, mean and median, of the associated relative pose [33], are used to evaluate the localisation error. We calculate ATE and RPE between the teach (mapping) trajectory and repeat (navigation) trajectory. Small ATE and RPE indicate that the robot can precisely navigate by following the previously taught path.
2) Navigation Robustness: the number of inliers in feature matching is used to measure the robustness of localisation. The visual offsets can reflect the stability of localisation. The number of inliers shows the confidence of localisation and a large number of inliers reflects that the robot localisation is robust and confident.

C. The Comparison and Baselines

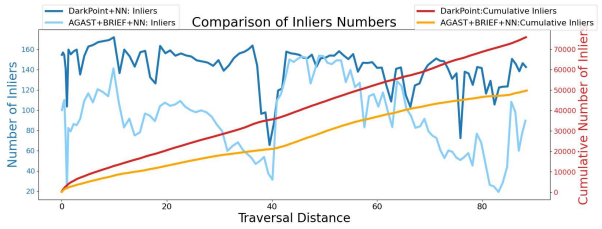
STROLL [2], [15], multi-experience map [4], [5], [6] and adaptive feature [10], [11] are the state-of-the-art visual teach-and-repeat navigation approaches. Among them, STROLL [15] is served as a baseline for comparison as it is the only open-sourced[4] monocular teach-and-repeat navigation system, to the best of our knowledge. As the visual feature matching is a core component of our system, we also assess the advances of the proposed DarkPoint descriptor. We integrated the following visual localisers with our system:

- STROLL-AGAST+BRIEF+NN: AGAST keypoint detector [34] and BRIEF descriptor [26] with nearest neighbour matching. This approach is widely used in state-of-the-art V-T&R methods [2], [15] or as base model of adaptive visual features [10], [11]. In our implementation, we use a maximum of 500 keypoints and a targeted number of 200 keypoints for the experiments.

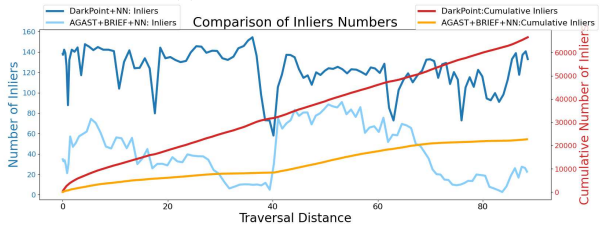
[4]Different implementations of the same algorithm may cause unobjective performance comparisons.



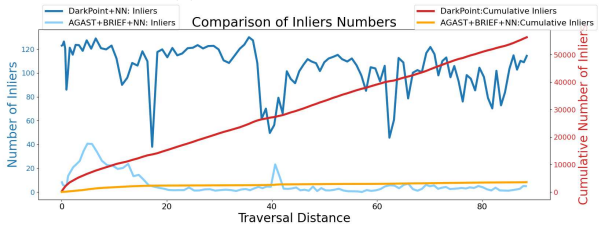
(a) Performed at 16:33.



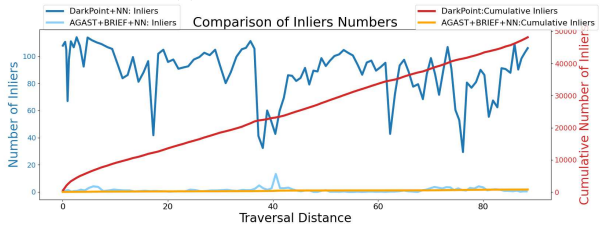
(b) Performed at 16.40.



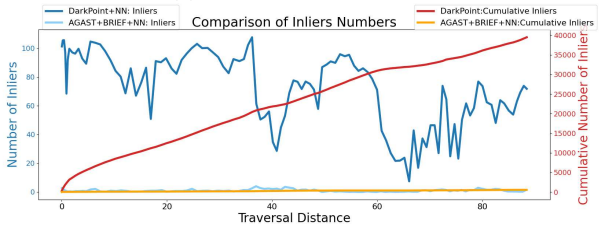
(c) Performed at 17:56.



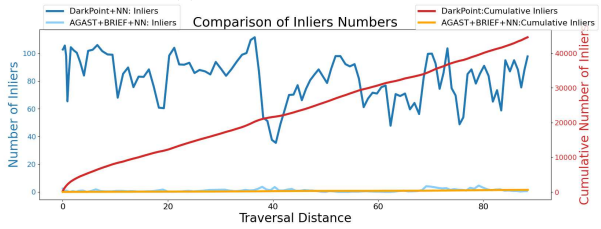
(d) Performed at 18:07.



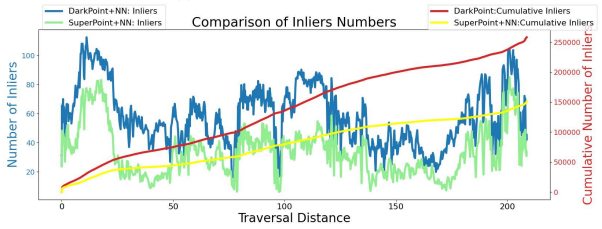
(e) Performed at 18:18.



(f) Performed at 18:33.



(g) STROLL performed at 18:56. Ours performed at 19:01.



(h) Church Test. Performed at 22:00.

Fig. 5: (a)-(g) are results of parking lot day-to-night experiments. Local sunset time when performing the experiment was 17:47. (h) is the result of the church experiment and local sunset was 18:51.

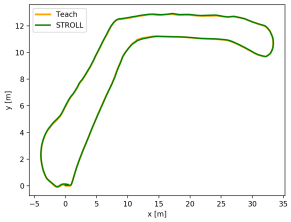
- Ours-DarkPoint+NN: DarkPoint with nearest neighbour matching. A maximum number of 500 keypoints, a detection threshold of 0.005 and a Non-Maximum Suppression radius of 4 pixels are used.
- Ours-SuperPoint+NN: SuperPoint with nearest neighbour matching. We use the pre-trained model from Magic Leap. The same settings are used with DarkPoint.
- Ours-SuperGlue: SuperPoint with Graph Neural Network matcher[31]. We use the pre-trained model from Magic Leap and the outdoor model is used.

Before running the long-term navigation experiments, we designed a multi-run experiments to verify the navigation repeatability of our system as well as the baseline system. We repeat the same trajectory for five times in constant lightness conditions. Both STROLL and ours are able to reproduce the quasi-identical navigation provided no significant change of the environment. The navigation trajectories and quantitative results are presented in Fig. 7 and Table II.

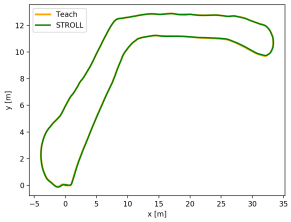
D. Day-to-Night Experiments (Parking-lot)

The aim of these experiments is to verify the robustness and advances of our V-T&R in terms of illumination changes and the impact of these changes on robot navigation accuracy. The experimental site was located at a parking-lot near Sheffield Robotics Centre, and the total navigation path was 86 m. In our experiments, we created the map at 16:00 (daytime) and evaluated the repeat navigation every 10 to 30 minutes for six rounds. We compare our system with the STROLL baseline [15]. The navigation robustness (i.e. feature matching) performance is illustrated in Fig. 5, the navigation trajectories are shown in Fig. 6 and the quantitative results are given in Table I.

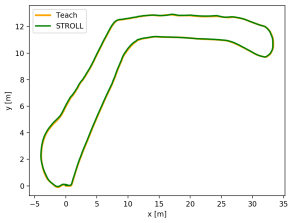
From Fig. 5, it can be seen that our method and STROLL produced similar amount of inliers in the teaching session. However, during the sunset, the inliers number of STROLL decreased dynamically and failed in the sixth repeat. In contrast, the number of inliers generated by our method are much more stable and remained at a relatively high level



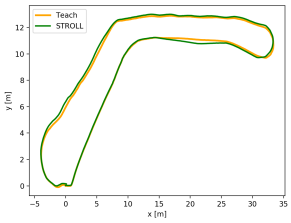
(1) Performed at 16:40.



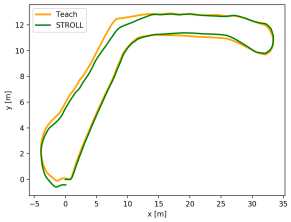
(2) Performed at 17:56.



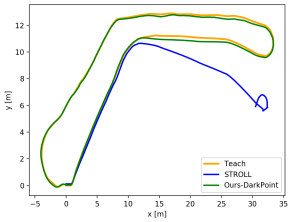
(3) Performed at 18:07.



(4) Performed at 18:18.



(5) Performed at 18:33.



(6) STROLL performed at 18:56. Ours-DarkPoint performed at 19:01.

Fig. 6: Parking lot day-to-night experiments. Local sunset time when performing the experiment was 17:47.

(above 50%) even after sunset. As shown in Table I and Fig. 6, the navigation accuracy of STROLL falls along with the decreasing feature matching robustness, but our system can navigate at night as accurate as STROLL in the daytime.

E. Slow Driving Experiment - Church

In this experiment, we investigate navigation performance at night using a daytime map by repeating a long trajectory under low illumination condition. This test was located at St. George's Church at the campus of the University of Sheffield, and the navigation route was 221 m with an average speed of 0.76m/s (shown in Fig. 8). The map was created in the afternoon at 15:00 and we tested the navigation at 22:00. In this experiment, we compared the navigation accuracy of our system with STROLL. STROLL failed in this experiment due to day-night illumination changes. The ATE and RPE are shown in Table I.

In particular, we also evaluated the navigation robustness (shown in Fig. 5), and compared to pretrained SuperPoint model, our tailored deep descriptor DarkPoint achieves approximately 1.7 times more inliers during the navigation. Hence the tailored DarkPoint shows better navigation robustness, compared to SuperPoint.

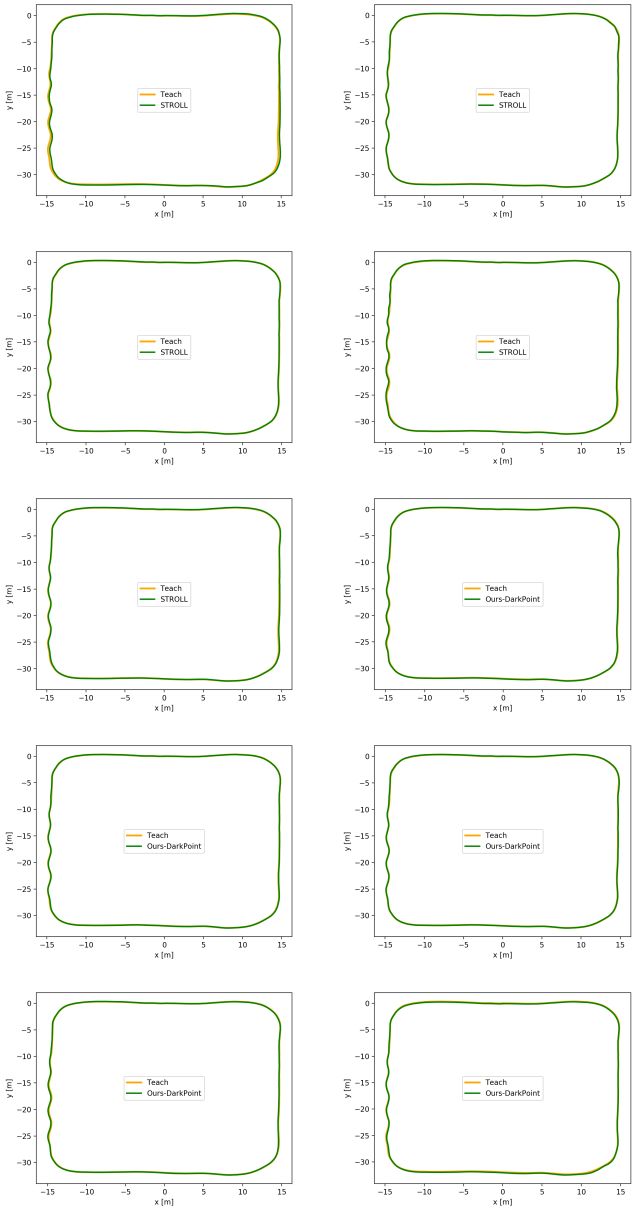


Fig. 7: The navigation repeatability experiments.

F. Long and Fast Navigation Experiment - Courtyard

In this experiment, we performed another day-to-night navigation experiment where the teaching phase took place in the afternoon and the repeat process took place in the evening with a path containing a mix of poorly lit segments and well artificially illuminated sections. The performance of STROLL, Ours-DarkPoint and Ours-SuperGlue was evaluated on a 434 m long trajectory, with an average speed of 1.47 m/s. As shown in Fig. 9, STROLL only managed to follow the trajectory for about a quarter of the trajectory then went off-road and the robot was manually forced to stop to avoid collisions. However our DarkPoint and SuperGlue methods successfully repeated the path. As shown by the data in table I. DarkPoint outperformed SuperGlue with about

Exp	Method	ATE(RMSE)	ATE(mean)	ATE(median)	RPE(RMSE)	RPE(mean)	RPE(median)	Result
Parking Lot 1	STROLL[15]	0.083	0.072	0.063	2.25	1.49	0.85	success
Parking Lot 2	STROLL[15]	0.074	0.064	0.054	2.06	1.43	0.93	success
Parking Lot 3	STROLL[15]	0.078	0.068	0.060	2.90	1.94	1.20	success
Parking Lot 4	STROLL[15]	0.148	0.130	0.115	3.36	2.36	1.52	success
Parking Lot 5	STROLL[15]	0.233	0.207	0.181	4.29	2.94	1.90	success
Parking Lot 6	STROLL[15]	0.830	0.671	0.530	13.58	6.36	1.82	fail
Parking Lot 6	Ours-DarkPoint	0.088	0.075	0.064	2.93	1.99	1.30	success
Church	STROLL[15]	1.07	1.01	0.93	19.05	10.00	2.30	fail
Church	Ours-DarkPoint	0.29	0.26	0.25	3.36	2.57	1.91	success
Courtyard	STROLL[15]	0.703	0.624	0.624	4.18	2.63	1.65	fail
Courtyard	Ours-DarkPoint	0.258	0.199	0.160	2.18	1.61	1.15	success
Courtyard	Ours-SuperGlue	0.391	0.302	0.236	4.50	3.02	1.90	success

TABLE I: Day-to-night navigation experiments. The unit of ATEs is in meters and unit of RPEs is in degrees.

Exp	Method	ATE(RMSE)	ATE(mean)	ATE(median)	RPE(RMSE)	RPE(mean)	RPE(median)	Result
Run #1	STROLL[15]	0.048	0.042	0.038	2.38	1.69	1.03	success
Run #2	STROLL[15]	0.056	0.051	0.048	3.05	2.14	1.34	success
Run #3	STROLL[15]	0.057	0.053	0.052	2.20	1.59	0.96	success
Run #4	STROLL[15]	0.042	0.034	0.028	3.44	2.43	1.60	success
Run #5	STROLL[15]	0.046	0.040	0.036	1.68	1.29	0.92	success
Run #6	Ours-DarkPoint	0.052	0.048	0.046	1.40	1.01	0.80	success
Run #7	Ours-DarkPoint	0.057	0.053	0.049	1.28	1.00	0.71	success
Run #8	Ours-DarkPoint	0.057	0.051	0.049	1.46	1.13	0.86	success
Run #9	Ours-DarkPoint	0.054	0.051	0.052	1.32	1.06	0.79	success
Run #10	Ours-DarkPoint	0.067	0.057	0.052	2.05	1.37	0.87	success

TABLE II: Multiple runs of each method under the same lighting conditions to measure performance consistency. The unit of ATEs is in meters and unit of RPEs is in degrees.

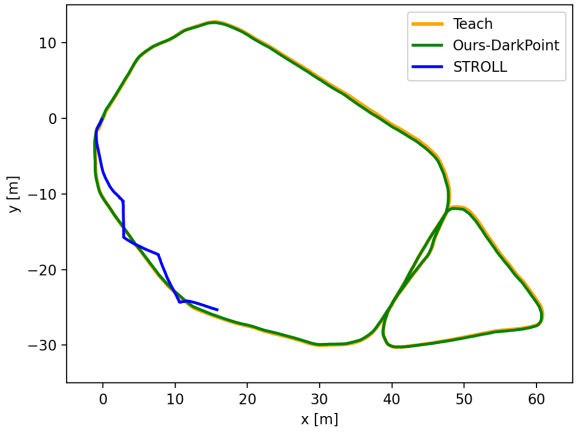


Fig. 8: The trajectory error in the church experiment.

half mean RPE and a lower mean ATE, demonstrating DarkPoint's lower drift over long-distance high-speed navigation.

SuperGlue is the state-of-the-art end-to-end matcher that uses SuperPoint as a front-end. The GNN can aggregate local features hence improve the matching performance. However, the bulky end-to-end matching lowers down the run-time performance to around 9-11Hz. V-T&R navigation requires a timely decision-making especially for repeating complex trajectories in high speed. This will be a trade-off between accuracy and efficiency. From our experience, the autonomous navigation is likely to fail when the localisation-control loop frequency drops lower than 7Hz. Hence, we can conclude that our system with DarkPoint and NN matching is superior than all comparison methods.

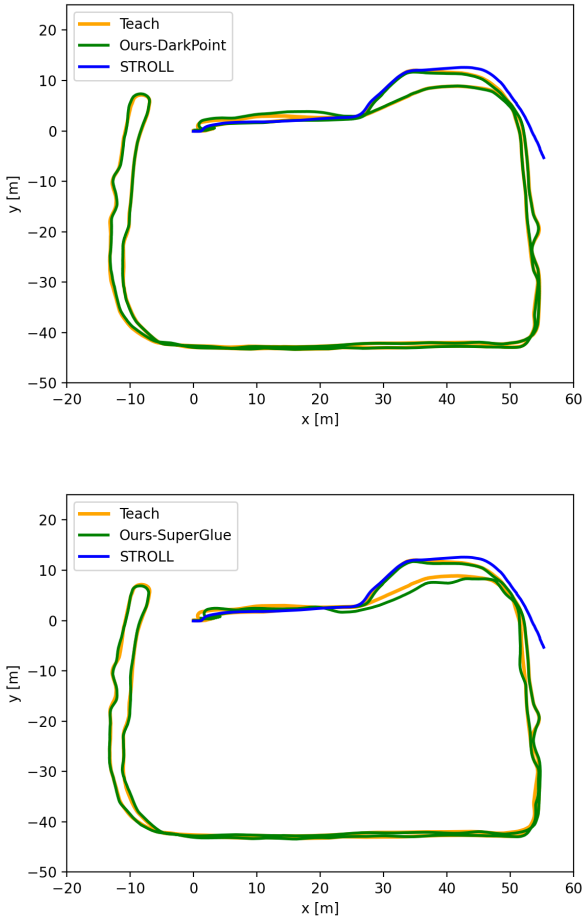


Fig. 9: Trajectories in the courtyard experiments.

V. Conclusion

In this paper, we proposed a robust monocular teach-and-repeat navigation system, where deep-learned descriptors are utilised to address the challenges of domain variance in long-term navigation. Specifically, the proposed approach is elastic and calibration free, and does not rely on precise metric mapping and explicit localisation. By fully leveraging advanced illumination adaptation in the local descriptor learning, our navigation system demonstrated day-to-night autonomous navigation using a single daytime map.

The experimental results show that the proposed navigation system is able to conduct a long-distance navigation task (more than 440 m, at an average speed of 1.47 m/s) in outdoor environments at night using a map created in the daytime with a very small trajectory error (0.25 m, 2.18°) achieved. The system is sufficiently robust to deal with high-speed manoeuvres and paths of complex shapes. Our system is open source and fully-integrated with ROS and the perception-action loop runs at 25-30Hz.

References

[1] P. Furgale and T. D. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, vol. 27, no. 5, pp. 534–560, 2010.

[2] T. Krajník, J. Faigl, V. Vonásek, K. Košnar, M. Kulich, and L. Přeučil, "Simple yet stable bearing-only navigation," *Journal of Field Robotics*, vol. 27, no. 5, pp. 511–533, 2010.

[3] M. Paton, K. MacTavish, L.-P. Berczi, S. K. van Es, and T. D. Barfoot, "I can see for miles and miles: An extended field test of visual teach and repeat 2.0," in *Field and Service Robotics*. Springer, 2018, pp. 415–431.

[4] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1645–1661, 2013.

[5] M. Paton, K. MacTavish, M. Warren, and T. D. Barfoot, "Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat," in *IROS*. IEEE, 2016, pp. 1918–1925.

[6] K. MacTavish, M. Paton, and T. D. Barfoot, "Visual triage: A bag-of-words experience selector for long-term visual route following," in *ICRA*. IEEE, 2017, pp. 2065–2072.

[7] P. Neubert, N. Sünderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *2013 European Conference on Mobile Robots*. IEEE, 2013, pp. 198–203.

[8] S. Lowry and M. J. Milford, "Supervised and unsupervised linear learning techniques for visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 32, no. 3, pp. 600–613, 2016.

[9] T. Krajník, P. Cristóforis, K. Kusumam, P. Neubert, and T. Duckett, "Image features for visual teach-and-repeat navigation in changing environments," *Robotics and Autonomous Systems*, vol. 88, pp. 127–141, 2017.

[10] N. Zhang, M. Warren, and T. D. Barfoot, "Learning place-and-time-dependent binary descriptors for long-term visual localization," in *ICRA*. IEEE, 2018, pp. 828–835.

[11] L. Halodová, E. Dvořráková, F. Majer, T. Vintr, O. M. Mozos, F. Dayoub, and T. Krajník, "Predictive and adaptive maps for long-term visual navigation in changing environments," in *IROS*. IEEE, 2019, pp. 7033–7039.

[12] L. Kunze, N. Hawes, T. Duckett, M. Hanheide, and T. Krajník, "Artificial intelligence for long-term robot autonomy: A survey," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4023–4030, 2018.

[13] K. Kidono, J. Miura, and Y. Shirai, "Autonomous visual navigation of a mobile robot using a human-guided experience," *Robotics and Autonomous Systems*, vol. 40, no. 2-3, pp. 121–130, 2002.

[14] E. Royer, M. Lhuillier, M. Dhome, and J.-M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 237–260, 2007.

[15] T. Krajník, F. Majer, L. Halodová, and T. Vintr, "Navigation without localisation: reliable teach and repeat based on the convergence theorem," in *IROS*. IEEE, 2018, pp. 1657–1664.

[16] S. Segvic, A. Remazeilles, A. Diosi, and F. Chaumette, "Large scale vision based navigation without an accurate global reconstruction," in *IEEE International Conference on Computer Vision and Pattern Recognition, CVPR'07*, Minneapolis, Minnesota, 2007, pp. 1–8.

[17] G. Blanc, Y. Mezouar, and P. Martinet, "Indoor navigation of a wheeled mobile robot along visual routes," in *ICRA*, 2005.

[18] Y. Matsumoto, M. Inaba, and H. Inoue, "Visual navigation using view-sequenced route representation," in *ICRA*, 1996.

[19] E. Royer, M. Lhuillier, M. Dhome, and J.-M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 237–260, Sep 2007.

[20] Z. Chen and S. T. Birchfield, "Qualitative vision-based path following," *IEEE Transactions on Robotics and Automation*, vol. 25, no. 3, pp. 749–754, 2009.

[21] T. Krajník, J. P. Fentanes, J. M. Santos, and T. Duckett, "Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments," *IEEE Trans. Robotics*, vol. 33, no. 4, pp. 964–977, 2017.

[22] F. Dayoub and T. Duckett, "An adaptive appearance-based map for long-term topological localization of mobile robots," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 3364–3369.

[23] D. M. Rosen, J. Mason, and J. J. Leonard, "Towards lifelong feature-based mapping in semi-static environments," in *ICRA*. IEEE, 2016, pp. 1063–1070.

[24] N. Carlevaris-Bianco and R. M. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2769–2776.

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

[26] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1281–1298, 2011.

[27] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint detection and description of local features," in *CVPR 2019*, 2019.

[28] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "Lf-net: learning local features from images," in *Advances in neural information processing systems*, 2018, pp. 6234–6244.

[29] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[30] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key. net: Keypoint detection by handcrafted and learned cnn filters," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5836–5844.

[31] 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*, 2020, pp. 4938–4947.

[32] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[33] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IROS*, Oct. 2012.

[34] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *European conference on Computer vision*. Springer, 2010, pp. 183–196.