

MVP: Unified Motion and Visual Self-Supervised Learning for Large-Scale Robotic Navigation

Marvin Chancán^{1,2} and Michael Milford¹

Abstract—Autonomous navigation emerges from both motion and local visual perception in real-world environments. However, most successful robotic motion estimation methods (e.g. VO, SLAM, SfM) and vision systems (e.g. CNN, visual place recognition–VPR) are often separately used for mapping and localization tasks. Conversely, recent reinforcement learning (RL) based methods for visual navigation rely on the quality of GPS data reception, which may not be reliable when directly using it as ground truth across multiple, month-spaced traversals in large environments. In this paper, we propose a novel motion and visual perception approach, dubbed *MVP*, that unifies these two sensor modalities for large-scale, target-driven navigation tasks. Our MVP-based method can learn faster, and is more accurate and robust to both extreme environmental changes and poor GPS data than corresponding vision-only navigation methods. MVP temporally incorporates compact image representations, obtained using VPR, with optimized motion estimation data, including but not limited to those from VO or optimized radar odometry (RO), to efficiently learn self-supervised navigation policies via RL. We evaluate our method on two large real-world datasets, Oxford Robotcar and Nordland Railway, over a range of weather (e.g. overcast, night, snow, sun, rain, clouds) and seasonal (e.g. winter, spring, fall, summer) conditions using the new *CityLearn* framework; an interactive environment for efficiently training navigation agents. Our experimental results, on traversals of the Oxford RobotCar dataset with no GPS data, show that MVP can achieve 53% and 93% navigation success rate using VO and RO, respectively, compared to 7% for a vision-only method. We additionally report a trade-off between the RL success rate and the motion estimation precision, suggesting that vision-only navigation systems can benefit from using precise motion estimation techniques to improve their overall performance.

I. INTRODUCTION

Navigation is a key component for enabling the deployment of mobile robots and autonomous vehicles in real-world environments. Current large-scale, real-world navigation systems rely on the usage of GPS data only as ground truth for sensory image labeling [1]–[6]. They then reduce the problem of navigation to vision-only methods [1], GPS-level localization combined with publicly available maps [2], or extend it with language-based tasks [3]–[6]. These end-to-end learning approaches are hard to train due to their large network models and weakly-related input sensor modalities.

The work of M.C. was supported by the Peruvian Government. The work of M.M. was supported by ARC grants FT140101229, CE140100016, the QUT Centre for Robotics, and the Australian Government via grant AUSMURIB000001 associated with ONR MURI grant N00014-19-1-2571.

¹ QUT Centre for Robotics, School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, Australia

² School of Mechatronics Engineering, Universidad Nacional de Ingeniería, Lima, Peru. mchancanl@uni.pe

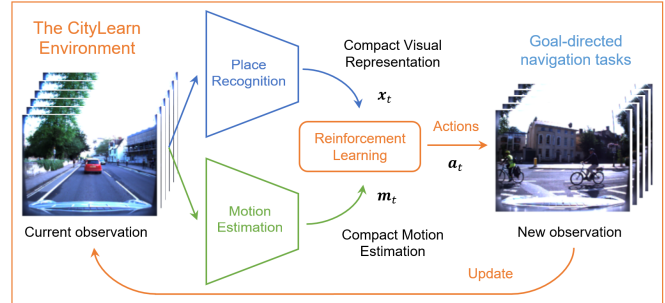


Fig. 1. **Proposed MVP-based approach.** We temporally incorporate odometry-based motion estimation data with compact image representations to perform large-scale all-weather navigation via reinforcement learning. Our method is efficient, accurate and robust to extreme environmental changes, even when GPS data reception fail (see Fig. 2).

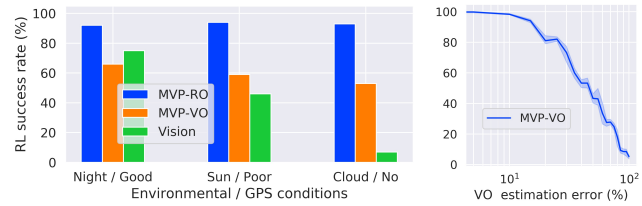


Fig. 2. **Navigation success rate results on the Oxford RobotCar dataset.** **Left:** Our MVP-based methods can accurately navigate across a range of visual environmental changes and GPS data situations, where vision-only approaches typically fail. **Right:** Trade-off curve of the RL success rate and the motion estimation precision using VO (log scale for better visualization).

Moreover, their generalization capabilities to environments with different visual conditions is not well explored. In contrast, we have recently shown an alternative non end-to-end vision-based approach using preprocessed compact image representations to achieve practical training and deployment on real data with challenging environmental transitions [7].

In this paper, we build on the main ideas of our previous work [7]—that combines reinforcement learning (RL) and visual place recognition (VPR) techniques for navigation tasks—to present a new, more efficient and robust approach. The main contributions of this paper are detailed as follows:

- Leveraging successful robotic motion estimation methods including VO [8] or radar [9] to capture compact motion information through an environment that can then be used to perform goal-driven navigation tasks (see Fig. 1). This makes our system more efficient and robust to extreme environmental changes, even with limited or no GPS data availability (Fig. 2-left).
- Using RL to temporally incorporate compact motion representations with equally compact image observations, obtained via deep-learning-based VPR models [10], for large-scale, all-weather navigation tasks.

- Experimental results on the RL navigation success rate and the VO motion estimation precision trade-off (Fig. 2-right). This shows how our proposed navigation system can improve its overall performance based on precise motion estimation techniques such as VO.

We evaluate our motion and visual perception (MVP) method using our interactive *CityLearn* framework [7], and present extensive experiments on two large real-world driving datasets, the Oxford RobotCar [11] and the Nordland Railway [12] datasets. The results of our MVP-based approach are consistently high across multiple, month-spaced traversals with extreme environmental changes, such as winter, spring, fall, and summer for Nordland, and overcast, night, snow, sun, rain, and clouds for Oxford (see blue bar in Fig. 2). For Nordland, we show how our approach outperforms corresponding vision-only navigation methods under extreme environmental changes, especially when GPS data is fully available and consistent across multiple traversals of the same route. For Oxford, we show the robustness of our approach across a range of real GPS data reception situations, including poor and no data reception at all, where vision-only navigation systems typically fail.

II. RELATED WORK

We present a brief overview of some successful work in motion estimation research, related visual place recognition methods for sequence-based driving datasets, and recent RL-based navigation systems for large real-world environments.

A. Motion Estimation in Robotics

Odometry-based sensors (i.e. wheel, inertial, laser, radar, and visual) for *self-localization* have long attracted attention in robotics research as an alternative approach to estimate motion information, especially in situations where GPS data is not reliable such as multi-path reception and changes in environmental conditions [8], [11], [13]. Traditional VO methods [14], SLAM-based systems [15], including MonoSLAM [16] and ORB-SLAM [17], and also bio-inspired models such as RatSLAM [18], [19] have captured the main challenges of the localization problem in large outdoor environments and indoor spaces—also with a range of alternative systems [20]–[23]. These methods have shown good invariance to changes in viewpoint and illumination by associating hand-crafted visual place features to locally optimized maps. Odometry, however, is known to be the first phase of solving a SLAM problem, which also includes loop closing and global map optimization. Consequently, multi-sensor fusion techniques combining vision [24]–[30], inertial sensors [31]–[33], LiDAR [34]–[36], and radar [37] have been proposed to further improve both the localization accuracy and the real-time performance, with a number of recent deep-learning-based methods that can match the precision of those traditional methods [9], [38]–[47]. In this work, we demonstrate our approach using both learning- and conventional-based odometry methods for motion estimation when GPS data is not available, as occurs in several traversals of the Oxford RobotCar dataset [11].

B. Visual Place Recognition

VPR approaches can be broadly split into two categories: image-retrieval-based methods that compare a single-frame (query) to an image database (reference) [10], [48]–[52], and multi-frame-based VPR techniques built on top of those single-frame methods to work on image sequences [53]–[57]; typically found in driving datasets [11], [12], [58]–[63]. These two approaches often first require computing image descriptors using diverse hand-crafted- [64] or deep-learning-based models [65]–[67]. However, using large image representations can be computationally expensive and also limit the deployment of these methods on real robots. Alternatively, we have recently demonstrated how compact image representations can be used to achieve state-of-the-art results in visual localization [68] by modeling temporal relationships between consecutive frames to improve the performance of compact single-frame-based methods. In this paper, we build on these main ideas to propose our MVP-based approach that uses compact but rich image representations, such as those from NetVLAD [10], and can also temporally use movement data through an environment via odometry-based techniques.

C. Learning-based Navigation

Significant progress has recently been made in goal-driven navigation tasks using learning methods [1]–[6], [69]–[80], inspired by advances in deep RL [81], [82]. Most of these algorithms can successfully train navigation agents end-to-end based on raw images. These approaches, however, are typically only evaluated using either synthetic data [69]–[72], [80], indoor spaces [73], [74] or relatively small outdoor environments [75], that generally do not require GPS data or map information. Alternatively, combining map-like- or SLAM-based input modalities, including motion sensor data, and images for goal-driven navigation tasks has been proposed [1], [76]–[79], but again these methods are trained only using small indoor environments. For large-scale outdoor navigation, however, different approaches that rely on GPS data as ground truth have been proposed [1], with a range of developments using language-based tasks [3]–[6] or publicly available maps [2]. However, relying on GPS data only for benchmarking purposes may not be reliable; especially when using large driving datasets recorded over many month-spaced traversals, as highlighted in previous work [9], [11].

In this paper, we propose a different approach that overcomes the limitations of prior work for large-scale, all-weather navigation tasks. We unify two fundamental and highly-related sensor modalities: motion and visual perception (MVP) information. Our MVP-based method builds on the main ideas presented in our previous works [7], [68]—that use compact image representations to achieve sample-efficient RL-based visual navigation using real data [7], and also demonstrate how to leverage motion information for VPR tasks [68]. We propose a network architecture that can incorporate motion information with visual observations via RL to perform accurate navigation tasks under extreme environmental changes and with limited or no GPS data; where visual-based navigation approaches typically fail. We

provide extensive experimental results in both visual place recognition and navigation tasks, using two large real-world dataset, that show how our method efficiently overcomes the limitations of those vision-only navigation pipelines.

III. MVP-BASED METHOD OVERVIEW

Our objective is to train an RL agent to perform goal-driven navigation tasks across a range of real-world environmental conditions, especially under poor GPS data conditions. We therefore developed an MVP-based approach that can be trained using motion estimation and visual data gathered in large environments. Our MVP method operates by temporally associating local estimates of motion with compact visual representations to efficiently train our policy network. Using this data, our policy can learn to associate motion representations with visual observations in a self-supervised manner, enabling our system to be robust to both visual changing conditions and poor GPS data.

In the following sections, we describe our problem formulation via RL, the driving datasets we used in our experiments, details of our MVP representations, our evaluation metrics for VPR and navigation tasks, and related visual navigation methods against which we compare our approach.

A. Problem Formulation

We formulate our goal-driven navigation tasks as a Markov Decision Process \mathcal{M} , with discrete state space $\mathbf{s}_t \in \mathcal{S}$, discrete action space $\mathbf{a}_t \in \mathcal{A}$, and transition operator $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ as in a finite-horizon T problem. Our goal is to find θ^* that maximizes this objective function:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}(\tau)} \left[\sum_{t=1}^T \gamma r(\tau) \right] \quad (1)$$

where $\pi_{\theta} : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ is the stochastic policy we want to learn, and $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function with discount factor γ . We parametrize our navigation policy π_{θ} with a neural network that can learn θ to optimize our policy. We also defined our state space \mathcal{S} by our compact bimodal MVP space representation $(\mathbf{m}_t, \mathbf{x}_t)$, and our action space \mathcal{A} by discrete action movements in the agent action space (\mathbf{a}_t) .

B. Real-World Driving Datasets

We evaluate our approach using our interactive CityLearn framework [7] on two challenging large real-world datasets, the Oxford RobotCar dataset [11] and the Nordland Railway dataset [12], that include diverse environmental changes and real GPS data reception situations.

Oxford RobotCar: This dataset [11] was collected using the Oxford RobotCar platform over a 10km route in central Oxford, UK. The data recorded with a range of sensors (e.g. LiDARs, monocular cameras and trinocular stereo cameras) includes more than 100 traversals (image sequences) of the same route with a large range of transitions across weather, season and dynamic urban environment changes over a period of 1 year. In Fig. 3 we show the selected 6 multiple traversals used in our experiments, referred here as *overcast*,



Fig. 3. Our six selected traversals from the Oxford RobotCar dataset.



Fig. 4. Samples of the four traversals provided by the Nordland dataset.

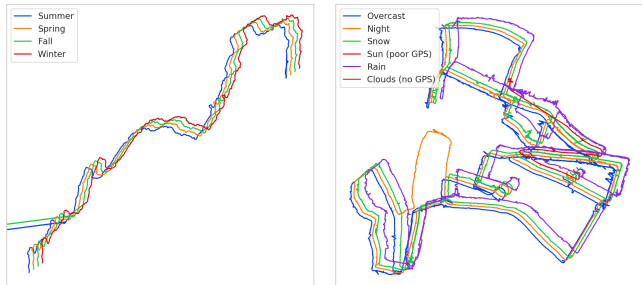


Fig. 5. **Raw GPS data:** The 4 traversals from the Nordland Railway (left) dataset and our 6 selected traversals from the Oxford RobotCar (right) dataset, both with a number of weather/season changes and GPS data reception situations (drifted for better visualization).

night, *snow*, *sun*, *rain*, and *clouds*.¹ Fig. 5-right shows the raw GPS data of our selected traversals, where both the *sun*, and the *clouds* traversals have poor GPS data reception and no GPS data at all, respectively.

Nordland Railway: The Nordland dataset [12] covers a 728km train journey from Trondheim to Bodø in Nordland, Norway. This 10 hour train ride has been recorded four times, once per season: *summer*, *spring*, *fall*, and *winter*. Fig. 4 shows a sample image for each traversal we used in our experiments, and Fig. 5-left shows the related raw GPS data; which is more consistent compared to the Oxford RobotCar raw GPS data (see Fig. 5-right).

C. Motion Estimation

To provide our agent with motion data we separately use three different sensor modalities in our experiments: raw GPS data, visual odometry (VO), and optimized radar odometry (RO). For the Oxford RobotCar dataset, it already provides both GPS and VO sensor data. For RO, we used the optimized ground truth RO sensor data provided in the extended Oxford Radar RobotCar dataset [9]—which has been demonstrated to be more accurate under challenging visual transitions—carefully chosen to visually match our selected traversals. For the Nordland dataset, however, we used the provided raw GPS data only as it is consistent across every traversal (Fig. 5-left).

The goal of using these two datasets is to evaluate the effectiveness of our MVP-based approach in situations where vision-only navigation methods typically fail. For Nordland, when GPS data is fully available and consistent, we show

¹Referred as 2015-05-19-14-06-38, 2014-12-10-18-10-50, 2015-02-03-08-45-10, 2015-08-04-09-12-27, 2015-10-29-12-18-17, and 2015-05-15-08-00-49, respectively, in [11].

that our approach can generalize better than visual-based navigation systems under extreme visual transitions. Similarly, for Oxford RobotCar, our method outperforms these visual-based navigation pipelines again under extreme visual changes, and also when GPS data reception fails.

D. Visual Representations

To enable sample-efficient RL training, as per previous work [7], we encode all our full resolution RGB sensory images using the off-the-shelf VPR model NetVLAD; based on a VGG-16 network architecture [83] with PCA plus whitening on top their model. This deep-learning-based model is known to provide significantly better image representations compared to other VPR approaches [84], and also enables to obtain compact feature dimensions (e.g., from 4096- d all the way down to 64- d). However, other deep-learning- or VPR-based models can equally used to encode our raw images. In this work, we used 64- d image representations, \mathbf{x}_t , in all our MVP-based experiments. We then combine it with compact 2- d motion representations, \mathbf{m}_t , to generate equally compact bimodal representations, \mathbf{b}_t , that feed our navigation policy network, see Fig. 6 (a). We encoded \mathbf{m}_t into compact feature vectors to preserve the compactness of \mathbf{b}_t , but it can be encoded using larger representations as in [1].

E. MVP-based Policy Learning

Goal-driven navigation: Our method is trained on both motion (\mathbf{m}_t) and visual representations (\mathbf{x}_t) to successfully navigate through actions (\mathbf{a}_t) towards a required goal destination (\mathbf{g}_t), which is also encoded using 2- d feature representations, over a single traversal in our CityLearn environment; see Fig. 1 and Fig. 6 (a) for further details.

Network Architecture: We design our network model inspired by [1], see Fig. 6 (a). A single *linear* layer with 512 units encodes our MVP bimodal representation (\mathbf{b}_t) to then combine it with the agent’s previous actions (\mathbf{a}_{t-1}), using a single recurrent layer long short-term memory (LSTM) [85] with 256 units. Updated agent’s actions (\mathbf{a}_t) are also used to estimate both the required next actions and the value function V from our policy network (π_θ). To optimize π_θ , we use the proximal policy optimization (PPO) algorithm [86], which evaluates our objective function in Eq. (1) for policy learning. We choose PPO as it can properly balance the small sample complexity of our compact input modalities and fine tuning requirements.

Reward design and curriculum learning: We use multiple levels of curriculum learning [87] to gradually encourage our agent to explore the environment, and a sparse reward function that gives the agent a reward of +1 only when it finds the target.

F. Vision-based Navigation Agent

We compare our MVP-based agent against a visual navigation agent with network architecture as proposed in [1], see Figs. 6 (a) and (b). This raw image based agent is adapted to also use a 2- d feature vector for \mathbf{g}_t to enable a fair comparison. In contrast to our method, this agent relies

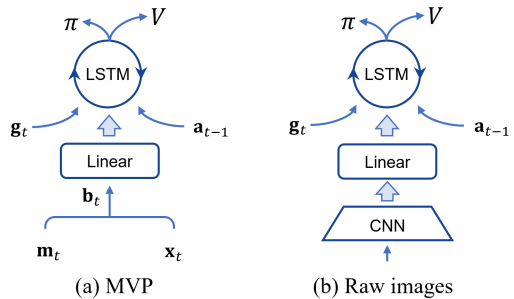


Fig. 6. **Baselines agents.** We compared (a) our MVP-based approach to an (b) end-to-end vision-based policy network model based on raw images and also relies on GPS data for ground truth labeling as in [1].

on GPS data for image labeling during both training and deployment and also does not incorporate motion learning. The network architecture of this agent includes a visual module of 2 *convolutional* layers, as per previous work [70], [88], with RGB input images of 84×84 pixels. The first CNN layer uses a kernel of size 8×8 , stride of 4×4 , and 16 feature maps, and the second CNN layer uses a kernel of size 4×4 , stride of 2×2 , and 32 feature maps.

G. Evaluation Metrics

VPR experiments: We report extensive VPR results, obtained using our compact image representations (\mathbf{x}_t), in order to provide an indicator of the visual component performance underlying our overall RL-based MVP system. A *linear* classifier is trained on each reference traversal to then evaluate it on the remaining query traversals. Classification scores obtained for each image are then used to compute precision-recall curves, which are finally used to calculate our *area under the curve* (AUC) results. AUC results on 10 experiments per traversal are presented in Fig. 7.

RL-based navigation tasks: We evaluate our trained agents on all corresponding dataset traversals, and provide statistics on the number of successful tasks in terms of the *success rate* results over 10 deployment iterations, each iteration with 100 different targets. Average results on those evaluations are reported in Fig. 8. We additionally constrain the maximum number of agent steps per episode to be less than the number of images within the traversal as in [7].

IV. EXPERIMENTS: RESULTS

We present two main experimental results: conventional single-frame visual place recognition evaluations (Fig. 7), and reinforcement learning deployment of the navigation agents; both evaluated in our two datasets (Fig. 8). We also provide details on the influence of incorporating motion data into our network architecture, and present selected illustrative results from our MVP method during deployment.

A. Visual Place Recognition Results

In Fig. 7 we provide a full overview of our visual place recognition experiments, as described in Sections III-D and III-G, in terms of AUC performance. For the Nordland dataset (Fig. 7-left), we trained a *linear* classifier on the

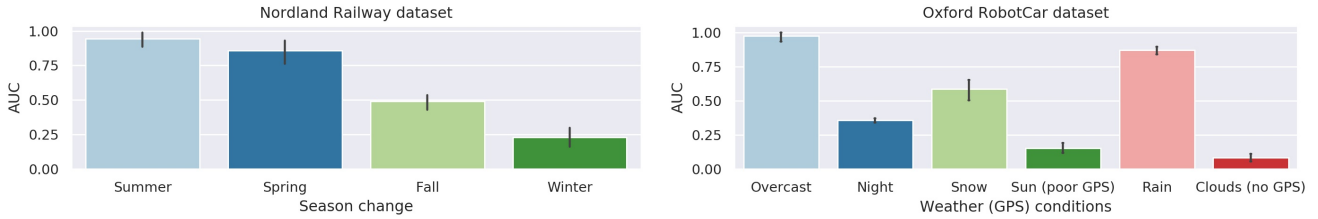


Fig. 7. **Visual place recognition experiments.** Conventional single-frame VPR results on Nordland Railway (left) and Oxford RobotCar (right) datasets. For Nordland, we show how our 64- d visual representations perform under season changes. For Oxford Robotcar, we additionally show how VPR methods suffer when using poor GPS data as ground truth.

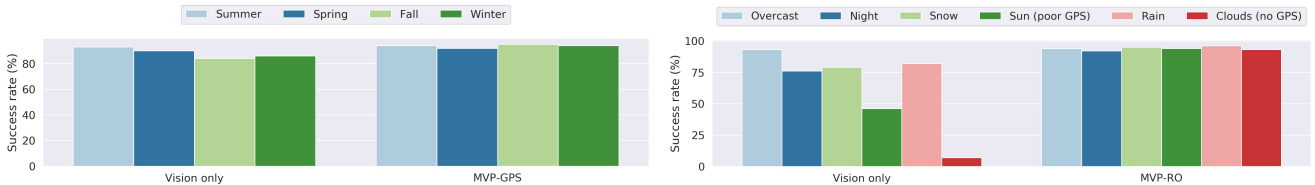


Fig. 8. **Reinforcement learning deployment.** Our MVP-based self-supervised method, that temporally aligns our compact visual representation with motion estimation data, achieves a 92% navigation success rate across all-weather conditions on the Nordland (left) and the Oxford RobotCar (right) dataset, compared to vision-only navigation methods that do not generalize well under extreme visual conditions (left) or rely on precise GPS data (right).

summer traversal, and then evaluated it on *spring*, *fall* and *winter* conditions. It is observed that extreme environmental changes such as those from *fall* and *winter* significantly reduce the results to around 0.50 and 0.25 AUC, respectively. It is worth noting, again, that each traversal of this dataset (and also of the Oxford RobotCar dataset) is a single sequence of images, meaning that we have used a single image from a particular place for training, and we do not use any data augmentation technique.

For the Oxford RobotCar dataset (Fig. 7-right), we trained a *linear* classifier on the *overcast* traversal, and evaluated it on the remaining traversals. AUC results in this dataset are relatively lower compared to those from the Nordland dataset; except for the *rain* traversal that achieves around 0.80 AUC. This is mainly because those traversals include significant viewpoint changes, diverse environmental transitions, and also real GPS data. For the *sun* and *clouds* traversals, with around 0.12 and 0.05 AUC, with poor and no GPS data reception, respectively, the results obtained highlight the importance of having good GPS data for ground truth labeling; especially for VPR. In contrast, the *night* and *snow* traversals, with good GPS data, still present relatively good result around 0.30 and 0.55 AUC, respectively; regardless of their significantly different visual and lighting conditions, compared to the *overcast* traversal.

B. Navigation Policy Deployment

Navigation success rate results of our RL policies are reported on the Nordland (Fig. 8-left) and Oxford RobotCar (Fig. 8-right) datasets. We also compare our method to a vision-only approach, as described in Sections III-E and III-F, respectively, using our CityLearn environment.

For the vision-based agent (referred as *vision only* in Fig. 8), which has been trained on raw images only, it is notable that the *success rate* results are significantly better than those

from our VPR experiments (Fig. 7), with over 84% success rate for the Nordland dataset (Fig. 8-left) and more than 46% for the Oxford RobotCar dataset (Fig. 8-right); except for the *clouds* traversal which has no GPS data. Suggesting that the generalization capabilities of the whole RL-based systems is robust to environmental variations. However, this method still does not generalize well under different weather conditions with significant viewpoint changes and occlusions, as in the Oxford RobotCar dataset, especially with poor GPS data (see *vision only* in 8-right for the *sun* and *clouds* traversals). Suggesting that RL-based vision-only navigation methods that rely on precise GPS information are likely to fail when using poor motion estimation information.

In contrast, our MVP-based agents overcome the limitations of the VPR module, underlying the vision-only method, by temporally incorporating those visual representations with precise motion information into our navigation policy network using either GPS (when fully available) or odometry-based techniques, referred in Fig. 8 as *MVP-GPS* and *MVP-RO*, respectively. On the Nordland dataset (Fig. 8-left), the *vision only* agent achieves around 86% success rate under challenging *winter* conditions, compared to around 94% for the MVP-GPS agent (see green bar in both cases). Similarly, on the Oxford RobotCar dataset (Fig. 8-right), the MVP-RO agent achieves 93% success rate under *clouds* conditions, with no GPS data available, compared to 7% for the *vision only* agent (see red bar in both cases).

C. Influence of Motion Estimation Precision

To analyze the influence of including motion data as an input to our policy network, we provide additional results on the Oxford RobotCar dataset shown in Fig. 2. Vision-based navigation methods actually generalize relatively well under extreme changes with a 75% success rate from day to night transitions, but with good GPS data reception.

However, these methods can fail when GPS data is not precise, even under similar visual conditions, such as day to clouds (day) changes, with a 7% success rate (see green bars for both cases in Fig. 2-left). Conversely, our MVP-based approach leverages the use of relatively precise motion estimation data, including but not limited to those from radar or visual odometry, on top of those vision navigation methods to improve overall performance under both visual changes and when there is no GPS data available (see orange and blue bars in Fig. 2-left). In Fig. 2-right, we characterize the deployment performance of our MVP-based method using VO to estimate motion data. This graph shows how incorporating precise motion information can improve the overall navigation performance of our system, suggesting that current vision-only navigation methods can also benefit from using MVP-like approaches. As also demonstrated in related work [14], odometry-based techniques can be used directly for navigation tasks. This method, however, may require additional baseline metrics to estimate global scale factors during deployment on real robots; particularly when using relative motion data relative to the robot initial pose.

D. Generalization Results

We present illustrative navigation deployment results in Figs. 9 and 10 for the vision-based agent and for our MVP-based approach, respectively. The agent is required to navigate from the same starting location to a distant target over all our selected traversals of the Oxford RobotCar dataset; see navigation states from left to right in Figs. 9 and 10 including two intermediate states. Our approach is capable of precisely navigating to the target for every condition change (see Fig. 10), while the vision-based agent fails under extreme condition variations and also where GPS data is poor or not available (Fig. 9).

V. CONCLUSIONS

We have proposed a method including a new network architecture that temporally integrate two fundamental sensor modalities such as motion and visual perception (MVP) information for large-scale target-driven navigation tasks using real data via reinforcement learning (RL). Our MVP-based approach was demonstrated to be robust to both extreme visual changing conditions and also poor absolute positioning information such as those from GPS, where typical visual (only) navigation pipelines fail. This suggests that the incorporation of motion information, including but not limited to GPS (when fully available) or visual/radar odometry, can be used to improve the overall performance and robustness of conventional visual-based navigation systems that rely on raw images only for learning complex navigation tasks. Future work combining different motion estimation modalities such as linear/angular velocities with visual representations is likely to be considered. However, this could potentially increase the network complexity and training requirements [80], [89], especially when using real data. Quantifying the relationship between required RL performance, visual place recognition generalization capabilities, and motion

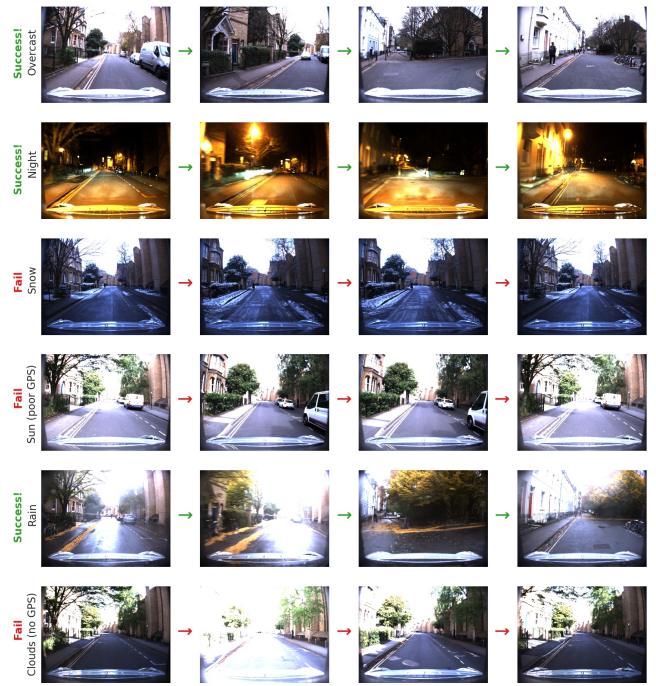


Fig. 9. **Vision-based navigation results.** This approach can navigate to the target on *overcast*, *night* and *rain*, but fails on the other traversals.

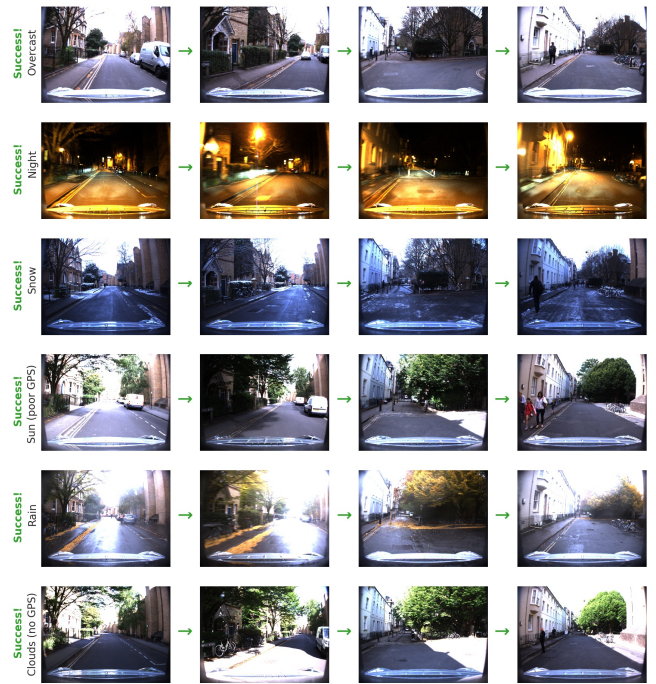


Fig. 10. **MVP-based navigation results.** Our approach is capable of precisely navigating to the goal across all the traversals of the Oxford RobotCar dataset.

estimation quality can also provide new insights for selecting between different motion estimation sensor modalities for a specific robotic navigation system.

REFERENCES

- [1] P. Mirowski *et al.*, “Learning to navigate in cities without a map,” in *Advances in Neural Information Processing Systems*, 2018, pp. 2419–2430.
- [2] H. Ma *et al.*, “Towards navigation without precise localization: Weakly supervised learning of goal-directed navigation cost map,” *arXiv preprint arXiv:1906.02468*, 2019.
- [3] V. Cirik *et al.*, “Following formulaic map instructions in a street simulation environment,” in *2018 NeurIPS Workshop on Visually Grounded Interaction and Language*, vol. 1, 2018.
- [4] K. M. Hermann, M. Malinowski, P. Mirowski, A. Banki-Horvath, K. Anderson, and R. Hadsell, “Learning to follow directions in street view,” *arXiv preprint arXiv:1903.00401*, 2019.
- [5] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, “Touchdown: Natural language navigation and spatial reasoning in visual street environments,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [6] H. de Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, and D. Kiela, “Talk the walk: Navigating new york city through grounded dialogue,” *arXiv preprint arXiv:1807.03367*, 2018.
- [7] M. Chancán and M. Milford, “From Visual Place Recognition to Navigation: Learning Sample-Efficient Control Policies across Diverse Real World Environments,” *arXiv preprint arXiv:1910.04335*, 2019, accepted to ICRA 2020.
- [8] S. A. S. Mohamed, M. Haghbayan, T. Westerlund, J. Heikkonen, H. Tenhunen, and J. Plosila, “A survey on odometry for autonomous navigation systems,” *IEEE Access*, vol. 7, pp. 97 466–97 486, 2019.
- [9] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, “The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset,” *arXiv preprint arXiv:1909.01300*, 2019.
- [10] R. Arandjelović, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.
- [11] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The oxford robotcar dataset,” *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [12] N. Sünderhauf, P. Neubert, and P. Protzel, “Are we there yet? challenging seqslam on a 3000 km journey across all four seasons,” in *Proc. Workshop Long-Term Autonomy 2013 IEEE Int. Conf. Robot. Autom. (ICRA)*, 2013.
- [13] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, “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, Dec 2016.
- [14] D. Nistér, O. Naroditsky, and J. Bergen, “Visual odometry,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2004.
- [15] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, “A solution to the simultaneous localization and map building (SLAM) problem,” *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229–241, June 2001.
- [16] 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, June 2007.
- [17] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM: A Versatile and Accurate Monocular SLAM System,” *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.
- [18] M. J. Milford, G. F. Wyeth, and D. Prasser, “Ratslam: a hippocampal model for simultaneous localization and mapping,” in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 1, April 2004, pp. 403–408.
- [19] M. J. Milford and G. F. Wyeth, “Mapping a suburb with a single camera using a biologically inspired slam system,” *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1038–1053, Oct 2008.
- [20] J. Biswas and M. Veloso, “Wifi localization and navigation for autonomous indoor mobile robots,” in *2010 IEEE International Conference on Robotics and Automation*, May 2010, pp. 4379–4384.
- [21] R. Miyagusuku, A. Yamashita, and H. Asama, “Improving gaussian processes based mapping of wireless signals using path loss models,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 4610–4615.
- [22] R. Miyagusuku, Y. Seow, A. Yamashita, and H. Asama, “Fast and robust localization using laser rangefinder and wifi data,” in *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Nov 2017, pp. 111–117.
- [23] R. Miyagusuku, A. Yamashita, and H. Asama, “Data Information Fusion From Multiple Access Points for WiFi-Based Self-localization,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 269–276, April 2019.
- [24] R. Mur-Artal and J. D. Tardós, “ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras,” *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, Oct 2017.
- [25] R. Mur-Artal and J. D. Tardós, “Visual-Inertial Monocular SLAM With Map Reuse,” *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, April 2017.
- [26] D. Nistér, O. Naroditsky, and J. Bergen, “Visual odometry for ground vehicle applications,” *Journal of Field Robotics*, vol. 23, no. 1, pp. 3–20, 2006.
- [27] N. L. Doh, H. Choset, and W. K. Chung, “Relative localization using path odometry information,” *Autonomous Robots*, vol. 21, no. 2, pp. 143–154, 2006.
- [28] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman, “Nidslam: Robust monocular slam using normalised information distance,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1435–1444.
- [29] P. Kim, B. Coltin, and H. J. Kim, “Low-drift visual odometry in structured environments by decoupling rotational and translational motion,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 7247–7253.
- [30] H. Zhan, C. S. Weerasekera, J. Bian, and I. Reid, “Visual odometry revisited: What should be learnt?” *arXiv preprint arXiv:1909.09803*, 2019.
- [31] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, “SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems,” *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, April 2017.
- [32] E. Hong and J. Lim, “Visual inertial odometry using coupled non-linear optimization,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 6879–6885.
- [33] B. P. W. Babu, D. Cyganski, J. Duckworth, and S. Kim, “Detection and resolution of motion conflict in visual inertial odometry,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 996–1002.
- [34] 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)*, May 2015, pp. 2174–2181.
- [35] P. Egger, P. V. K. Borges, G. Catt, A. Pfrunder, R. Siegwart, and R. Dub, “Posemap: Lifelong, multi-environment 3d lidar localization,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 3430–3437.
- [36] Z. Wang, J. Zhang, S. Chen, C. Yuan, J. Zhang, and J. Zhang, “Robust high accuracy visual-inertial-laser slam system,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 6636–6641.
- [37] S. H. Cen and P. Newman, “Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 6045–6052.
- [38] S. Wang, R. Clark, H. Wen, and N. Trigoni, “Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2043–2050.
- [39] S. Pillai and J. J. Leonard, “Towards visual ego-motion learning in robots,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 5533–5540.
- [40] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [41] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, “Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 340–349.
- [42] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Unsupervised monocular depth and ego-motion learning with structure and se-

- mantics,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [43] X. Wang *et al.*, “Improving learning-based ego-motion estimation with homomorphism-based losses and drift correction,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 970–976.
- [44] S. Y. Loo, A. J. Amiri, S. Mashohor, S. H. Tang, and H. Zhang, “Cnn-svo: Improving the mapping in semi-direct visual odometry using single-image depth prediction,” in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 5218–5223.
- [45] T. Shen, Z. Luo, L. Zhou, H. Deng, R. Zhang, T. Fang, and L. Quan, “Beyond photometric loss for self-supervised ego-motion estimation,” in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 6359–6365.
- [46] T. Shen, L. Zhou, Z. Luo, Y. Yao, S. Li, J. Zhang, T. Fang, and L. Quan, “Self-supervised learning of depth and motion under photometric inconsistency,” in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- [47] Ş. Săftescu, M. Gadd, D. De Martini, D. Barnes, and P. Newman, “Kidnapped radar: Topological radar localisation using rotationally-invariant metric learning,” *arXiv preprint arXiv:2001.09438*, 2020.
- [48] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, “24/7 place recognition by view synthesis,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [49] T. Weyand, I. Kostrikov, and J. Philbin, “Planet-photo geolocation with convolutional neural networks,” in *European Conference on Computer Vision*. Springer, 2016, pp. 37–55.
- [50] H. J. Kim, E. Dunn, and J. Frahm, “Learned contextual feature reweighting for image geo-localization,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3251–3260.
- [51] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, “End-to-end learning of deep visual representations for image retrieval,” *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 2017.
- [52] F. Radenović, G. Toliás, and O. Chum, “Fine-tuning cnn image retrieval with no human annotation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, July 2019.
- [53] M. J. Milford and G. F. Wyeth, “SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights,” in *2012 IEEE International Conference on Robotics and Automation (ICRA)*, May 2012, pp. 1643–1649.
- [54] E. Pepperell, P. I. Corke, and M. J. Milford, “All-environment visual place recognition with SMART,” in *2014 IEEE international conference on robotics and automation (ICRA)*, 2014, pp. 1612–1618.
- [55] N. Sünderhauf *et al.*, “Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free,” *Proceedings of Robotics: Science and Systems XII*, 2015.
- [56] N. Sünderhauf *et al.*, “On the performance of convnet features for place recognition,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 4297–4304.
- [57] S. Hausler, A. Jacobson, and M. Milford, “Multi-process fusion: Visual place recognition using multiple image processing methods,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1924–1931, 2019.
- [58] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3354–3361.
- [59] J. Guo, U. Kurup, and M. Shah, “Is it safe to drive? an overview of factors, challenges, and datasets for driveability assessment in autonomous driving,” *arXiv preprint arXiv:1811.11277*, 2018.
- [60] T. Naseer, W. Burgard, and C. Stachniss, “Robust visual localization across seasons,” *IEEE Transactions on Robotics*, vol. 34, no. 2, pp. 289–302, April 2018.
- [61] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, “FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day,” in *2010 IEEE International Conference on Robotics and Automation*, May 2010, pp. 3507–3512.
- [62] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [63] H. Caesar *et al.*, “nuScenes: A multimodal dataset for autonomous driving,” *arXiv preprint arXiv:1903.11027*, 2019.
- [64] S. Lowry *et al.*, “Visual place recognition: A survey,” *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, Feb 2016.
- [65] Z. Chen *et al.*, “Convolutional neural network-based place recognition,” *arXiv preprint arXiv:1411.1509*, 2014.
- [66] Z. Chen *et al.*, “Deep learning features at scale for visual place recognition,” *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3223–3230, 2017.
- [67] S. Garg, N. Sünderhauf, and M. Milford, “Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics,” *Proceedings of Robotics: Science and Systems XIV*, 2018.
- [68] M. Chancán, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, “A hybrid compact neural architecture for visual place recognition,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 993–1000, April 2020.
- [69] G. Kahn, A. Villafior, P. Abbeel, and S. Levine, “Composable action-conditioned predictors: Flexible off-policy learning for robot navigation,” in *Proceedings of the 2nd Annual Conference on Robot Learning*, 2018, pp. 806–816.
- [70] Piotr Mirowski *et al.*, “Learning to navigate in complex environments,” *arXiv preprint arXiv:1611.03673*, 2016.
- [71] D. S. Chaplot, E. Parisotto, and R. Salakhutdinov, “Active neural localization,” *arXiv preprint arXiv:1801.08214*, 2018.
- [72] J. Zhang, L. Tai, J. Boedecker, W. Burgard, and M. Liu, “Neural slam: Learning to explore with external memory,” *arXiv preprint arXiv:1706.09520*, 2017.
- [73] G. Kahn, A. Villafior, B. Ding, P. Abbeel, and S. Levine, “Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 5129–5136.
- [74] C. Oh and A. Cavallaro, “Learning action representations for self-supervised visual exploration,” in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 5873–5879.
- [75] G. Kahn, P. Abbeel, and S. Levine, “BADGR: An Autonomous Self-Supervised Learning-Based Navigation System,” *arXiv preprint arXiv:2002.05700*, 2020.
- [76] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, “Cognitive mapping and planning for visual navigation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2616–2625.
- [77] S. Gupta, D. Fouhey, S. Levine, and J. Malik, “Unifying map and landmark based representations for visual navigation,” *arXiv preprint arXiv:1712.08125*, 2017.
- [78] T. Chen, S. Gupta, and A. Gupta, “Learning exploration policies for navigation,” *arXiv preprint arXiv:1903.01959*, 2019.
- [79] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, “Learning to explore using active neural slam,” in *International Conference on Learning Representations*, 2020.
- [80] A. Banino *et al.*, “Vector-based navigation using grid-like representations in artificial agents,” *Nature*, vol. 557, no. 7705, pp. 429–433, 2018.
- [81] X. Guo, S. Singh, H. Lee, R. L. Lewis, and X. Wang, “Deep learning for real-time atari game play using offline monte-carlo tree search planning,” in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014, pp. 3338–3346.
- [82] V. Mnih *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [83] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [84] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. D. McDonald-Maier, “Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions,” *arXiv preprint arXiv:1903.09107*, 2019.
- [85] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
- [86] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [87] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 41–48.
- [88] L. Espeholt *et al.*, “IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures,” *arXiv preprint arXiv:1802.01561*, 2018.
- [89] C. J. Cueva and X.-X. Wei, “Emergence of grid-like representations by training recurrent neural networks to perform spatial localization,” *arXiv preprint arXiv:1803.07770*, 2018.