Self-Supervised, Goal-Conditioned Policies for Navigation in Unstructured Environments

Travis Manderson*, Juan Camilo Gamboa*, Stefan Wapnick*, Jean-François Tremblay*,
Hanqing Zhao*, Florian Shkurti†, Dave Meger* and Gregory Dudek*
*Mobile Robotics Laboratory, School of Computer Science, McGill University, Montreal, Canada
†Robot Vision & Learning Lab, Department of Computer Science, University of Toronto, Canada
Email: {travism,gamboa,swapnick,jft,hzhao,dmeger,dudek}@cim.mcgill.ca, florian@cs.toronto.edu

***Abstract*—We present a goal-conditioned visual navigation system trained using hindsight relabelling and self-supervision that learns a control policy for close-proximity robot inspection in natural, rugged environments. Our technique enables robots to navigate collision free to sparse geographic waypoints provided by the user without any prior map of the environment.**

We first learn a safe policy that greedily navigates to examine regions of interest while avoiding collisions but does not take into account specific geometric goals. We then augment this policy via the addition of goal-conditioning to seek our specific waypoints and a final goal.

The first policy synthesis is achieved through either behavior cloning or self-supervised learning. This policy is extended to be goal-conditioned, using hindsight relabelling, guided by the robot's relative localization system without any manual annotation. The goal-conditioned policy can then be deployed to navigate a series of arbitrary goals (waypoints).

We validated our approach on an underwater vehicle in a difficult open ocean environment to collect scientifically relevant data on coral reefs while operating safely and autonomously at an impressive proximity of 0.5 m to the coral. We also demonstrate this approach to terrestrial navigation, illustrated using a 1:5 scale, off-road vehicle.

I. Introduction

In this paper, we present a visual navigation system that learns safe, reactive behaviours to make close-proximity robot inspections in rugged and challenging environments. Our approach uses a combination of self-supervised reinforcement learning and hindsight relabelling to synthesize collision avoidance, informative path planning, and goal-directed navigation into a single policy that enables a robot to reach relative geometric waypoints provided by a user without any prior map. Our system is generic enough to be used in multiple unstructured domains, for example, underwater or on-ground, to autonomously collect scientific data critical for environmental monitoring and understanding.

Our approach begins by learning a non-goal-conditioned but safe navigation policy that seeks preferable terrain characteristics or scientifically desirable observations. In our previous work, we have shown that this policy can be trained via self-supervised reinforcement learning using a hybrid model-based and model-free network [25] or through behavior cloning where a relatively small set of images are labeled with steering commands that would direct the robot towards regions of interest [23, 24]. Using these policies, the robot is able to navigate safely following desired behaviour, but it is unable to reach specific geometric points.

Fig. 1. Navigation example for both an underwater and ground vehicle. Examples of waypoints to be visited are illustrated by the red circles (although in practice they are usually much more widely separated). The actual trajectory used to achieve them in the presence of obstacles is at the discretion of the learned policy.

We address this shortcoming in our proposed method by augmenting this policy with goal-seeking behaviour using hindsight experience relabelling [11]. We generate a dataset of images, actions, and position (as measured by a vision-based state estimator running onboard) collected from exploration experience using the non-goal-conditioned policy. We split the recorded trajectories into sub-segments and interpret the end position of each sub-segment as a goal waypoint. In this way, each segment is automatically labelled for goal-directed navigation from the beginning to end. We then use these goal-directed segments to train a goal-conditioned navigation policy that achieves safe, well behaved, and goal-directed navigation.

We have demonstrated our system in an underwater ocean environment with difficult visual conditions and external disturbances (Fig. 1 left), as well as performed preliminary validation on a 1:5 scale off-road vehicle (Fig. 1 right).

II. Background and Related Work

Our work builds upon existing literature on sensor-based navigation, imitation learning, and reinforcement learning for autonomous robot navigation in the field. Natural and unstructured environments, such as underwater or forests, often present unique navigation challenges, including incomplete knowledge of the surroundings, reacting to environmental disturbances, identifying free space, and inferring traversable terrain.

A common visual navigation approach in these environments is to first perform semantic segmentation followed by geometric path planning [10, 38, 41, 43]. Even with segmentation, often assigning a weight corresponding to the

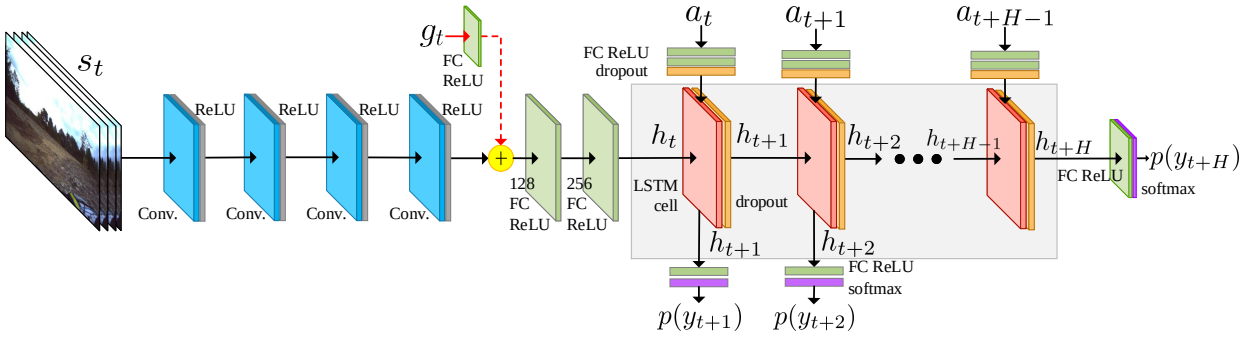

Fig. 2. Simplified model overview. Initially the model is trained using self supervision to predict terrain roughness and collision for a given image state and a set of steering action sequences. Following data collection of the initial model, a goal-conditioned model (with the goal added, as depicted with the red dashed arrow) is trained with an augmented dataset using hindsight relabelling. The goal-conditioned model can be successfully used to simultaneously traverse smooth terrain, avoid obstacles, and reach a set of desired waypoints.

traversability of each segmentation class is difficult. One approach is to infer the traversable classes using expert demonstrations [4, 42]

Imitation Learning: Another technique of relying on expert demonstrations is to directly predict actions from input images. Simple imitation can be done via supervised learning (behavioural cloning) [33] on observation and action pairs from an expert; however, it is not robust to distribution shift. Several authors have proposed data augmentation or domain matching methods for increasing robustness [7, 14, 15, 19], but the underlying problem of compounding errors during distribution shift remains. DAgger [37] addresses this problem by iteratively querying the expert on states that are visited by the learned policy during evaluation. In addition to ground vehicles, this technique has also been demonstrated on both aerial vehicles [36] and underwater robot systems [23]. More recent methods have combined human demonstrations and imitation learning [40] with end-to-end reinforcement learning [20] as another way to handle distribution shift.

Several authors have also used imitation learning as a way of identifying natural paths in the environment [34]. Demonstrations are often collected with one camera directed directly towards the path and several other cameras mounted with offset positions [14, 22]. In this way, the images captured can be automatically labelled with the corresponding action a robot should take to steer towards the path.

Reinforcement Learning: In reinforcement learning, a robot learns a navigation policy by maximizing the expected reward for a set of state-actions pairs. These methods can generally be classified into model-free or model-based, each having their own advantages and drawbacks. Oh *et al.* [30] recently proposed Value Prediction Networks (VPNs), which learn an encoded abstract state optimized to predict immediate rewards and value functions. The intuition is that planning need not require the full observation image. Instead, the prediction can be done on a learned model of encoded abstract states. The authors assert that VPNs can be viewed as jointly model-based and model-free and found them to be more sample efficient and to generalize better than Deep Q-Networks [28].

Khan *et al.* [17] developed a similar architecture, known as Generalized Computation Graphs (GCGs), for the task of predicting obstacles for a short horizon and demonstrated it on a remote-controlled car in an indoor hallway environment. We adopted a similar approach with the extension of predicting terrain roughness for off-road driving in unstructured environments [25].

Goal-Conditioned Navigation: While typical navigation learning scenarios consider a single task, such as lane following or navigating towards a fixed goal, we want also want to consider goal-conditioned policies. Conditional imitation learning has been studied extensively in the last few years [9, 11, 21, 32, 35] and has many connections to goal-conditioned reinforcement learning [2, 16, 39], particularly in the batch case.

Self-Supervised Learning: The idea of learning terrain classes or navigation strategies directly from sensor data has been examined by several authors. Giguere *et al.* [13] and [12, 26] looked at learning terrain classification and gait policy selection directly from unsupervised or minimally-labelled Inertial Measurement Unit (IMU) data. Wellhausen *et al.* [44] used recorded force-torque signals from a quadruped robot to label traversed terrain. This data was projected into the robot's camera frame for generating a training set for segmentation.

Our work is close in motivation to informative path planning [3, 6, 8, 27, 29, 31], where most methods need to estimate the surrounding map while executing frontier-based exploration. In contrast, our work does not assume a map because it does not need to perform exploration exploitation within a map due to the direct use of reactive policies instead of a pipeline consisting of perception, mapping, path planning, and tracking.

III. Approach

Our system is designed with the purpose of extracting a goal-conditioned policy, useful for navigation, from a lower level policy trained via self-supervised learning. The purpose of our design is that high-level behaviours, such as relevant scientific data collection or waypoint following, is harmoniously built on top of low-level behaviors like obstacle avoidance. Although our method is flexible enough to work with behavioural

cloned policies, for the purposes of this section, we focus on the model used for self-supervised learning for off-road navigation. Details on our goal-conditioned behavioral model can be found in [24].

A. Model Overview

We first start with a non-goal-conditioned deep learning model for predicting the terrain roughness and collision probabilities over a fixed horizon. Our architecture is similar to Khan *et al.* [17], which was previously shown to produce good performance for short-term control situations. The model architecture, as exemplified in Fig. 2, operates on the recent visual history of the past four images, representing the image state s_t and an action sequence of steering commands $\langle a_t, a_{t+1}, ...a_{t+H-1} \rangle$. The image state is passed through a convolution layer to form the initial hidden state h_t. of the Long Short-Term Memory (LSTM).

At each timestep, the model predicts the probability of each terrain class over the planning horizon (H): $\langle p(y_{t+1}), p(y_{t+2}), ..., p(y_{t+H}) \rangle$, where $y_t \in C$ and C is the set of all terrain classes. We choose increasing label values to be rougher terrain (0 being completely smooth and $|C|-1$ being an obstacle). This architecture models the joint probability of terrain classes over the horizon while assuming conditional independence between labels:

$$p(y_{t+H}, ...y_{t+1} | a_{t+H-1} ... a_t, s_t) = \prod_{i=1}^{H} p(y_{t+i} | a_{t+i-1} ... a_t, s_t)$$

The cross-entropy loss is used to train the model with L2 regularization:

$$L_t = -\sum_{i=1}^{H} \sum_{c_j \in C} \mathbf{1}_{y_{t+i}=c_j} \log p(\hat{y}_{t+i} | a_{t+i-1} ... a_t, s_t) + \lambda \|w\|_2^2$$

where y_{t+i} and $\hat{y}_{t+i}$ are the true and predicted labels.

B. Planning

Planning is performed by maximizing the expected reward over the planning horizon. Each terrain class is assigned a reward, with smooth terrain having the highest reward and the obstacle class having a reward of zero:

$$A_t^H = \underset{a_t, ...a_{t+H-1}}{\arg\max} - \sum_{i=1}^{H} \sum_{c_j \in C} c_j * p(\hat{y}_{t+i} = c_j | a_{t+i-1} ... a_t, s_t)$$

A randomized K-shooting method is used to generate a random set of K action rollouts (trajectory) at each timestep: $A_{(t,k)}^H = (a_t, ...a_{t+H-1})_k$. The expected cumulative reward for each trajectory is the sum of the individual rewards for each predicted terrain class. Our policy applies the first action from the trajectory with the highest perceived reward. Planning is repeated at every timestep, in an MPC fashion, to compensate for modelling errors.

C. System Overview

Our experimental vehicle used for off-road driving is a 1:5 scale RC buggy, as shown in Fig. 3. An embedded microcontroller is used to control a drive motor and two servo motors are used for steering and braking. The microcontroller also performs pose estimation using a Kalman filter to fuse

Fig. 3. Off-road vehicle used for real-world experiments. In practice, only one forward-facing camera is used in this work.

information from several redundant IMUs and an RTK GPS. An Intel i7 NUC running Ubuntu and Robot Operating System (ROS) records camera images, Lidar obstacle detections, and IMU readings. An NVIDIA Jetson Xavier, also running ROS, runs our deep learning model implemented in Tensorflow [1]. The Lidar is used only to perform short-range collision detection and act as a *bump* sensor and measure the instantaneous presence of an obstacle.

The IMU is used to measure terrain roughness, which has previously been considered by other authors [5, 18, 45]. We measure the Root Mean Square (RMS) linear acceleration reported by the robot's IMU over short time windows of 20 samples collected at 60 Hz to approximate an instantaneous roughness score of the traversed terrain.

D. Hindsight Relabelling

To train the goal-conditioned model, we generate a location-aware dataset by allowing the robot to explore using the non-goal-conditioned terrain model while recording the relative pose onboard. These trajectories are broken up into examples used for the navigation task: *which action should the robot execute to arrive at a particular location?*

Goals are sampled from the collected trajectories in a method similar to [2]. For each training mini-batch, a random set of trajectories and timesteps within those trajectories are selected from the dataset. Each sampled item is randomly associated with a future timestep (limited to ensure the goal is not excessively far away), which is interpreted as a goal. After association with a goal from hindsight relabelling, training tuples that can be applied to the goal-conditioned network are in the format $\langle \mathbf{s}_i, \mathbf{g}_i, \hat{\mathbf{a}}_i \rangle$, where $\mathbf{s}_i$ is the image state, $\mathbf{g}_i$ is the relative goal, and $\hat{\mathbf{a}}_i$ is the action taken.

E. Goal Conditioning

Our proposed goal-conditioned navigation policy is implemented by augmenting the non-goal-conditioned terrain prediction model with an input waypoint g_t relative to the robot's current frame. The input vector is passed through a fully-connected layer before being multiplied with the output of the convolution layer. In practice, we assume our method can be used when there is no access to global position coordinates. Therefore, we use onboard state estimation (e.g. IMU, GPS or visual odometry) to compute the relative offset to the waypoints that are with respect to the robot's starting position.

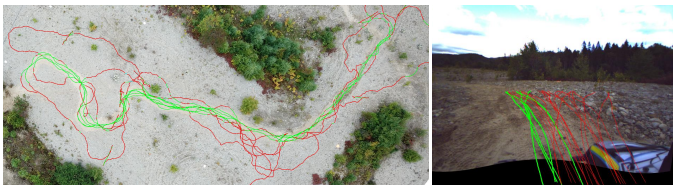

Fig. 4. Left: Example path executed during training in the real-world, overlaid on an aerial image. Measured smooth terrain is shown in green, while rough terrain is shown in red. Right: Sample of prediction trajectories to remain on the smooth terrain. High reward predictions are shown in green while low reward predictions are shown in red.

The model objective function is modified to predict the distance between the input action sequence $\langle a_t, a_{t+1}, ...a_{t+H-1} \rangle$ and the record action sequence $\langle \hat{a}_t, \hat{a}_{t+1}, ...\hat{a}_{t+H-1} \rangle$. In this way, we reward predictions with the lowest distance. We perform planning in a similar fashion to the non-goal-conditioned model by choosing the trajectory with the highest perceived cumulative reward and take the first action in the sequence.

IV. Experiments

Off-road Driving: We validated the self-supervised learning of terrain classes and the non-goal-conditioned model in the real-world, as shown in Fig. 4, on the 1:5 scale RC buggy described in section III-C. Despite a relatively small dataset, we obtained reasonable performance after ≈8,000 training steps. The prediction accuracy ranged from 78% to 60% over horizon lengths of 1 to 16 respectively. We also performed a quantitative evaluation of our model in simulation for four terrain classes: smooth, medium, rough, and collision. As shown in Table I, on-policy navigation significantly reduced the amount of rough terrain encountered from 42% to 11% and drove on smooth terrain 80% of the time.

TABLE I
Percentage of terrain traversed in simulation while driving on-policy and random.

	smooth	medium	rough	collision
on-policy	80.13%	8.47%	11.22%	0.18%
random	32.86%	22.42%	42.27%	2.45%

We performed a preliminary validation of our goal-conditioned model in simulation for off-road driving. We created training trajectories and sampled successive waypoints from a normal distribution with a mean of ≈28 m and a standard derivation of ≈13 m.

Fig. 5 illustrates the predicted action sequence from the goal-conditioned model that guides the robot towards the goal

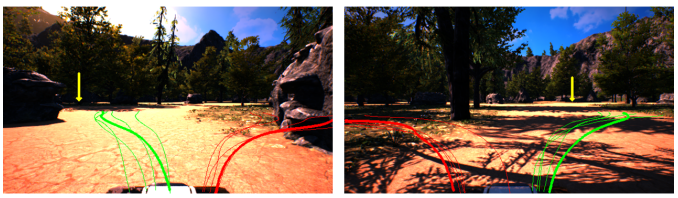

Fig. 5. Illustration of predicted trajectories towards a goal (marked by the yellow arrow), predictions with high and low values (bold lines being the most extreme) are shown in green and red colors respectively.

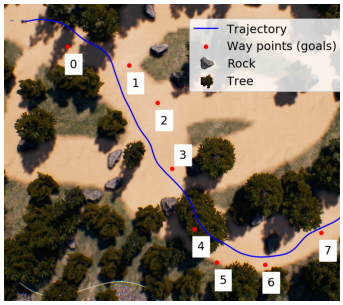

Fig. 6. Example trajectory with 8 waypoints traversed using the goal-conditioned policy. We consider a waypoint reached when it's within a short distance of the waypoint to allow the model to deviate from the direct path to maintain obstacle avoidance and drive on smooth terrain.

while still inheriting the obstacle avoidance and smooth terrain preferring behavior from the non-goal-conditioned model. Fig. 6 shows is an example trajectory traversed by the robot to reach a set of 8 waypoints while maintaining these characteristics.

Underwater Navigation: We have also successfully deployed a variant of the goal-conditioned policy in the open ocean on the underwater vehicle shown in Fig. 1 (left). In these experiments, we initially trained a behavioural imitation model described in [24] to avoid obstacles and prefer swimming over coral reef for scientific data sampling. Fig. 7 shows one trajectory that was executed to reach a set of waypoints. Overall, we ran trials executing 4 different 10-waypoint trajectories collectively spanning over 1 km. One of these trajectories reached 8 out of 10 waypoints, two reached 7 out of 10 waypoints, and one reached all 10 waypoints.

V. Conclusion

In this paper, we have described a self-supervised approach to learn goal-conditioned navigation policies in unstructured and natural environments. Additionally, we have examined its application in both off-road driving in rugged environments and in demanding underwater environments where the presence of surge, weather variations, and lighting variations cannot be overstated.

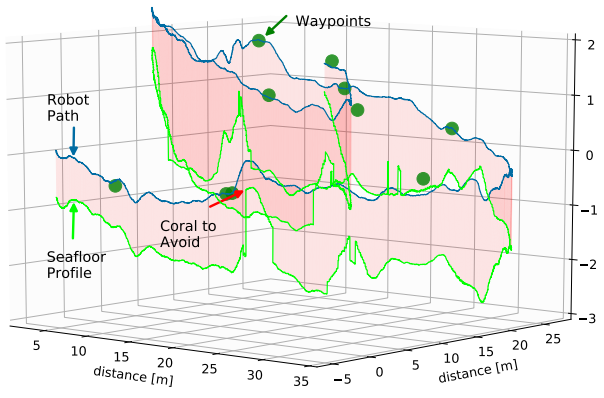

Fig. 7. Example trajectory of an underwater robot in the ocean navigating through several waypoints. The robot travels towards the waypoints while also avoiding obstacles as reflected by the change in the robot path with respect to the seafloor profile.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5048–5058. Curran Associates, Inc., 2017.

[3] Akash Arora, P. Michael Furlong, Robert Fitch, Salah Sukkarieh, and Terrence Fong. Multi-modal active perception for information gathering in science missions. *Computing Research Repository*, 2017.

[4] Dan Barnes, Will Maddern, and Ingmar Posner. Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 203–210, 2017. ISBN 9781509046331. doi: 10.1109/ICRA.2017.7989025.

[5] A Barsi, T Lovas, I Kertész, et al. The potential of lowend imus for mobile mapping systems. *International Archives of Photogrammetry and Remote Sensing*, 36(Part 1):4, 2006.

[6] J. Binney, A. Krause, and G. S. Sukhatme. Informative path planning for an autonomous underwater vehicle. In *IEEE International Conference on Robotics and Automation*, pages 4791–4796, May 2010.

[7] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *Computing Research Repository*, 2016.

[8] Benjamin Charrow, Gregory Kahn, Sachin Patil, Sikang Liu, Kenneth Y. Goldberg, Pieter Abbeel, Nathan Michael, and Vijay Kumar. Information-theoretic planning with trajectory optimization for dense 3d mapping. In *Robotics: Science and Systems*, 2015.

[9] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4693–4700, 2018.

[10] Jeffrey Delmerico, Elias Mueggler, Julia Nitsch, and Davide Scaramuzza. Active autonomous aerial exploration for ground robot path planning. *IEEE Robotics and Automation Letters*, 2(2):664–671, 2017. ISSN 23773766. doi: 10.1109/LRA.2017.2651163.

[11] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. In *Advances in Neural Information Processing Systems 32*, pages 15298–15309. 2019.

[12] E. M. DuPont, R. G. Roberts, and C. A. Moore. Speed independent terrain classification. In *2006 Proceeding of the Thirty-Eighth Southeastern Symposium on System Theory*, pages 240–244, 2006.

[13] Philippe Giguere, Gregory Dudek, Chris Prahacs, and Shane Saunderson. Environment identification for a running robot using inertial and actuator cues. In *Robotics: Science and Systems*, pages 271–278, 2006.

[14] A. Giusti, J. Guzzi, D. C. Cireşan, F. L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. D. Caro, D. Scaramuzza, and L. M. Gambardella. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1(2):661–667, July 2016. ISSN 2377-3766.

[15] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4565–4573. Curran Associates, Inc., 2016.

[16] Leslie Pack Kaelbling. Learning to achieve goals. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, Chambery, France, 1993. Morgan Kaufmann.

[17] Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. Self-Supervised Deep Reinforcement Learning with Generalized Computation Graphs for Robot Navigation. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 5129–5136, 2018. ISBN 9781538630815. doi: 10.1109/ICRA.2018.8460655.

[18] I Kertesz, Tamas Lovas, and Arpad Barsi. Measurement of road roughness by low-cost photogrammetric system. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36, 01 2007.

[19] Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 143–156. PMLR, 13–15 Nov 2017.

[20] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(1): 1334–1373, January 2016.

[21] X. Lin, P. Guo, C. Florensa, and D. Held. Adaptive variance for changing sparse-reward environments. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3210–3216, 2019.

[22] A. Loquercio, A. I. Maqueda, C. R. del Blanco, and D. Scaramuzza. Dronet: Learning to fly by driving. *IEEE Robotics and Automation Letters*, 3(2):1088–1095, April

2018.

[23] Travis Manderson, Juan Camilo Gamboa-Higuera, Ran Cheng, and Gregory Dudek. Vision-based autonomous underwater swimming in dense coral for combined collision avoidance and target selection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1885–1891. IEEE, 2018.

[24] Travis Manderson, Juan Camilo Gamboa Higuera, Stefan Wapnick, Jean-François Tremblay, Florian Shkurti, Dave Meger, and Gregory Dudek. Vision-based goal-conditioned policies for underwater navigation in the presence of obstacles. *Robotics: Science and Systems XVI*, 2020.

[25] Travis Manderson, Stefan Wapnick, Dave Meger, and Gregory Dudek. Learning to Drive Off-Road on Smooth Terrains in Unstructured Environments Using an Onboard Camera and Sparse Aerial Images. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation*, may 2020.

[26] Sandeep Manjanna, Gregory Dudek, and Philippe Giguere. Using gait change for terrain sensing by robots. In *2013 International Conference on Computer and Robot Vision*, pages 16–22. IEEE, 2013.

[27] Alexandra Meliou, Andreas Krause, Carlos Guestrin, and Joseph M. Hellerstein. Nonmyopic informative path planning in spatio-temporal models. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, page 602–607. AAAI Press, 2007.

[28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing Atari with Deep Reinforcement Learning. *CoRR*, abs/1312.5602, 2013.

[29] Haruki Nishimura and Mac Schwager. Sacbp : Belief space planning for continuous-time dynamical systems via stochastic sequential action control. 2019.

[30] Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6120–6130, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

[31] L. Paull, S. Saeedi, H. Li, and V. Myers. An information gain based adaptive path planning method for an autonomous underwater vehicle using sidescan sonar. In *2010 IEEE International Conference on Automation Science and Engineering*, pages 835–840, Aug 2010.

[32] M. Pfeiffer, M. Schaeuble, J. Nieto, R. Siegwart, and C. Cadena. From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1527–1533, 2017.

[33] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pages 305–313, 1989.

[34] Christopher Rasmussen, Yan Lu, and Mehmet Kocamaz. A trail-following robotwhich uses appearance and structural cues. In *Springer Tracts in Advanced Robotics*, volume 92, pages 265–279. Springer, Berlin, Heidelberg, 2014. ISBN 9783642406850. doi: 10.1007/978-3-642-40686-7_18.

[35] Nicholas Rhinehart, Rowan McAllister, and Sergey Levine. Deep imitative models for flexible inference, planning, and control. *arXiv preprint arXiv:1810.06544*, 2018.

[36] S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert. Learning monocular reactive uav control in cluttered natural environments. In *2013 IEEE International Conference on Robotics and Automation*, pages 1765–1772, 2013.

[37] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.

[38] Brandon Rothrock, Ryan Kennedy, Chris Cunningham, Jeremie Papon, Matthew Heverly, and Masahiro Ono. *SPOC: Deep Learning-based Terrain Classification for Mars Rover Missions*. doi: 10.2514/6.2016-5539.

[39] Tom Schaul, Dan Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1312–1320. JMLR.org, 2015.

[40] Avi Singh, Larry Yang, Kristian Hartikainen, Chelsea Finn, and Sergey Levine. End-to-end robotic reinforcement learning without reward engineering. *Robotics: Science and Systems*, 2019.

[41] Boris Sofman, Ellie Lin, J. Andrew Bagnell, John Cole, Nicolas Vandapel, and Anthony Stentz. Improving robot navigation through self-supervised online learning. *Journal of Field Robotics*, 23(11-12):1059–1075, 2006. doi: 10.1002/rob.20169.

[42] L. Tang, X. Ding, H. Yin, Y. Wang, and R. Xiong. From one to many: Unsupervised traversable area segmentation in off-road environment. In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 787–792, Dec 2017. doi: 10.1109/ROBIO.2017.8324513.

[43] Luqi Wang, Daqian Cheng, Fei Gao, Fengyu Cai, Jixin Guo, Mengxiang Lin, and Shaojie Shen. A collaborative aerial-ground robotic system for fast exploration, 2018.

[44] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter. Where Should I Walk? Predicting Terrain Properties From Images Via Self-Supervised Learning. *IEEE Robotics and Automation Letters*, 4(2):1509–1516, April 2019. ISSN 2377-3774. doi: 10.1109/LRA.2019.2895390.

[45] Wan Wen. Road roughness detection by analysing imu data. Master's thesis, Royal Institute of Technology (KTH), 2008.