PIRLNav: Pretraining with Imitation and RL Finetuning for OBJECTNAV

Anonymous ICLR submission

Paper ID 18

Abstract

We study ObjectGoal Navigation– where a virtual robot situated in a new environment is asked to navigate to an object. Prior work [1] has shown that imitation learning (IL) on a dataset of human demonstrations achieves promising results. However, this has limitations – 1) IL policies generalize poorly to new states, since the training mimics actions not their consequences, and 2) collecting demonstrations is expensive. On the other hand, reinforcement learning (RL) is trivially scalable, but requires careful reward engineering to achieve desirable behavior. We present a two-stage learning scheme for IL pretraining on human demonstrations followed by RL-finetuning. This leads to a PIRLNav policy that advances the state-of-the-art on OBJECTNAV *from* 60.0% *success rate to* 65.0% *(+*5.0% *absolute).*

Using this IL→RL training recipe, we present a rigorous empirical analysis of design choices. First, we investigate whether human demonstrations can be replaced with 'free' (automatically generated) sources of demonstrations, e.g. *shortest paths (SP) or task-agnostic frontier exploration (FE) trajectories. We find that IL→RL on human demonstrations outperforms IL→RL on SP and FE trajectories, even when controlled for the same IL-pretraining success on* TRAIN, *and even on a subset of* VAL *episodes where IL-pretraining success favors the SP or FE policies. Next, we study how RL-finetuning performance scales with the size of the IL pretraining dataset. We find that as we increase the size of the IL-pretraining dataset and get to high IL accuracies, the improvements from RL-finetuning are smaller, and that* 90% *of the performance of our best IL→RL policy can be achieved with less than half the number of IL demonstrations. Finally, we analyze failure modes of our* OBJECTNAV *policies, and present guidelines for further improving them.*

1. Introduction

Since the seminal work of Winograd [2], designing embodied agents that have a rich understanding of the environment they are situated in, can interact with humans (and other agents) via language, and the environment via actions has been a long-term goal in AI [3–12]. We focus on ObjectGoal Navigation [13, 14], wherein an agent situated in a new environment is asked to navigate to any instance of an object category ('find a plant', 'find a bed', *etc.*); see Fig. 5 (in Appendix). OBJECTNAV is simple to explain but difficult for today's techniques to accomplish. First, the agent needs to be able to ground the tokens in the language instruction to physical objects in the environment (*e.g.* what does a 'plant' look like?). Second, the agent needs to have rich semantic priors to guide its navigation to avoid wasteful exploration (*e.g.* the microwave is likely to be found in the kitchen, not the washroom). Finally, it has to keep track of where it has been in its internal memory to avoid redundant search.

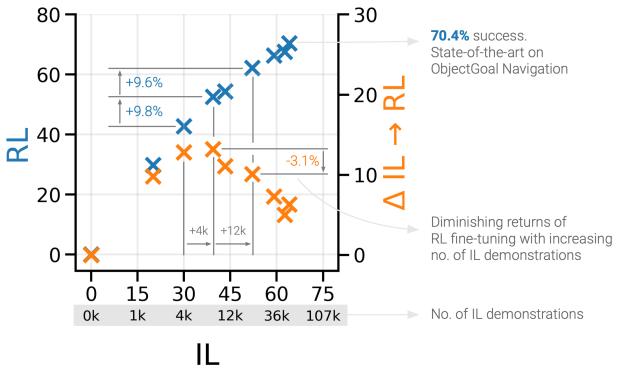

Figure 1. OBJECTNAV success rates of agents trained using imitation learning (IL) *vs*. IL-pretraining followed by reinforcement learning (RL) (in blue). RL from scratch (*i.e.* IL=0) fails to get off-the-ground. With more IL demonstrations, IL success increases, and it transfers to even higher RL-finetuning success. But the difference between RL-finetuning *vs*. IL-pretraining success (in orange) plateaus and starts to decrease beyond a certain point, indicating diminishing returns with each additional IL demonstration.

Humans are adept at OBJECTNAV. Prior work [1] collected a large-scale dataset of $80k$ human demonstrations for OBJECTNAV, where human subjects on Mechanical Turk teleoperated virtual robots and searched for objects in novel houses. This first provided a human baseline on OBJECTNAV of 88.9% success rate on the Matterport3D (MP3D)

dataset [15][1] compared to 35.4% success rate of the best performing method [1]. This dataset was then used to train agents via imitation learning (specifically, behavior cloning) of human demonstrations.

While this approach achieved state-of-art results (35.4% success rate on MP3D VAL dataset), it has two clear limitations. First, imitation learning (behavior cloning) is known to suffer from poor generalization to out-of-distribution states not seen during training, since the training emphasizes imitating actions not accomplishing their goals. Second and more importantly, it is expensive and thus not scalable. Specifically, Ramrakhya *et al*. [1] collected $80k$ demonstrations on 56 scenes in Matterport3D Dataset, which took $\sim$2894 hours of human teleoperation and \$$50k$ dollars. A few months after [1] was released, a new higher-quality dataset called HM3D-Semantics v0.1 [16] became available with 120 annotated 3D scenes, and a few months after that HM3D-Semantics v0.2 added 96 additional scenes. Scaling Ramrakhya *et al*.'s approach to continuously incorporate new scenes involves replicating that entire effort again and again. On the other hand, training with reinforcement learning (RL) is trivially scalable once annotated 3D scans are available. However, as demonstrated in Maksymets *et al*. [17], RL requires careful reward engineering, the reward function typically used for OBJECTNAV actually *penalizes* exploration (even though the task requires it), and the existing RL policies overfit to the small number of available environments.

Our primary technical contribution is PIRLNav, an approach for pretraining with IL and finetuning with RL for OBJECTNAV. IL pretrained policies provide a reasonable starting point for 'bootstrapping' RL and make the optimization easier than learning from scratch. In fact, we show that IL pretraining even unlocks RL with sparse rewards. Sparse rewards are simple (do not involve any reward engineering) and do not suffer from the unintended consequences described above. However, learning from scratch with sparse rewards is typically out of reach since most random action trajectories result in no positive rewards.

While combining IL and RL has been studied in prior work [18–22], the main technical challenge in the context of modern neural networks is that imitation pretraining results in weights for the policy (or actor), but not a value function (or critic). Thus, naively initializing a new RL policy with these IL-pretrained policy weights often leads to catastrophic failures due to destructive policy updates early on during RL training, especially for actor-critic RL methods [23]. To overcome this challenge, we present a two-stage learning scheme (in Appendix Sec. B) involving a critic-only learning phase first that gradually transitions over to training both the actor and critic. We also identify a set of practical recommendations for this recipe to be applied to OBJECTNAV. This leads to a PIRLNav policy that advances the state-the-art on OBJECTNAV from 60.0% success rate (in [24]) to 65.0% (+5.0%, 8.3% relative improvement).

Next, using this IL+RL training recipe, we conduct an empirical analysis of design choices. Specifically, an ingredient we investigate is whether human demonstrations can be replaced with 'free' (automatically generated) sources of demonstrations for OBJECTNAV, *e.g*. (1) shortest paths (SP) between the agent's start location and the closest object instance, or (2) task-agnostic frontier exploration [25] (FE) of the environment followed by shortest path to goal-object upon observing it. We ask and answer the following questions:

1. *'Do human demonstrations capture any unique* OBJECTNAV*-specific behaviors that shortest paths and frontier exploration trajectories do not?'* Yes. We find that IL / IL+RL on human demonstrations outperforms IL / IL+RL on shortest paths and frontier exploration trajectories respectively. When we control the number of demonstrations from each source such that IL success on TRAIN is the same, RL-finetuning when initialized from IL on human demonstrations still outperforms the other two.
2. *'How does performance after RL scale with IL dataset size?'* We observe diminishing returns from RL-finetuning as we scale IL dataset size. This suggests, by effectively leveraging the trade-off curve between size of pretraining dataset size *vs*. performance after RL-Finetuning, we can achieve closer to state-of-the-art results without investing into a large dataset of IL demonstrations.
3. *'Does IL on frontier exploration demonstrations present similar scaling behavior as IL on human demonstrations?'* No. We find that as we scale frontier exploration demonstrations past $70k$ trajectories, the performance plateaus.

Finally, we present an analysis of the failure modes of our OBJECTNAV policies (in Appendix Sec. C) and present a set of guidelines for further improving them. Our policy's primary failure modes are: a) Dataset issues: comprising of missing goal annotations, and navigation meshes blocking the path, b) Navigation errors: primarily failure to navigate between floors, c) Recognition failures: where the agent does not identify the goal object during an episode, or confuses the specified goal with a semantically-similar object.

2. Role of demonstrations in IL→RL transfer

Our decision to use human demonstrations for IL-pretraining before RL-finetuning was motivated by results in prior work [1]. Next, we examine if other cheaper sources of demonstrations lead to equally good IL→RL generalization. Specifically, we consider 3 sources of demonstrations:

Shortest paths (SP). These demonstrations are generated by greedily sampling actions to fit the geodesic shortest path to the nearest navigable goal object, computed using the ground-truth map of the environment. These demonstrations do not capture any exploration, they only capture success at the OBJECTNAV task via the most efficient path.

[1] On VAL split, for 21 object categories, and a maximum of 500 steps.

Training demonstrations	Success (↑)	SPL (↑)
Shortest paths ($240k$)	6.4%	5.0%
Frontier exploration ($70k$)	44.9%	21.5%
Human demonstrations ($77k$)	**64.1%**	**27.1%**

Table 1. Performance on HM3D VAL with imitation learning on SP, FE, and HD demonstrations. The size of each demonstration dataset is picked such that total steps of experience is $\sim 12M$.

Task-Agnostic Frontier Exploration (FE) [24]. These are generated by using a 2-stage approach: 1) Exploration: where a task-agnostic strategy is used to maximize exploration coverage and build a top-down semantic map of the environment, and 2) Goal navigation: once the goal object is detected by the semantic predictor, the developed map is used to reach it by following the shortest path. These demonstrations capture OBJECTNAV-agnostic exploration.

Human Demonstrations (HD) [1]. These are collected by asking humans on Mechanical Turk to control an agent and navigate to the goal object. Humans are provided access to the first-person RGB view of the agent and tasked to reach within 1m of the goal object category. These demonstrations capture human-like OBJECTNAV-specific exploration.

2.1. Results with Imitation Learning

Using the IL approach described in Appendix Sec. A.3, we train on SP, FE, and HD demonstrations. Since these demonstrations vary in trajectory length (*e.g.* SP are significantly shorter than FE), we collect $\sim 12M$ steps of experience with each method. That amounts to $240k$ SP, $70k$ FE, and $77k$ HD demonstrations respectively. Our IL results are summarized in Table 1. IL on $240k$ SP demonstrations leads to 6.4% success and 5.0% SPL. We believe this poor performance is due to an imitation gap [26], *i.e.* the shortest path demonstrations are generated with access to privileged information (ground-truth map of the environment) which is not available to the policy during training. Without a map, following the shortest path in a new environment to find a goal object is not possible. IL on $70k$ FE demonstrations achieves 44.9% success and 21.5% SPL, which is significantly better than IL on shortest paths (+38.5% success, +16.5% SPL). Finally, IL on $77k$ HD obtains the best results – 64.1% success and 27.1% SPL. These trends suggest that task-specific exploration (captured in human demonstrations) leads to much better generalization than task-agnostic exploration (FE) or shortest paths (SP).

2.2. Results with RL Finetuning

Using the IL-pretrained policies on SP, FE, and HD demonstrations as initialization, we RL-finetune each using our approach described in Appendix Sec. B. These results are summarized in Fig. 2. Perhaps intuitively, the trends after RL-finetuning follow the same ordering as IL-pretraining, *i.e.* RL-finetuning from IL on HD > FE > SP. But there are

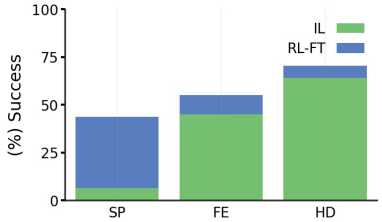

Figure 2. OBJECTNAV performance on HM3D VAL with IL-pretraining on shortest path (SP), frontier exploration (FE), and human demonstrations (HD), followed by RL-finetuning from each.

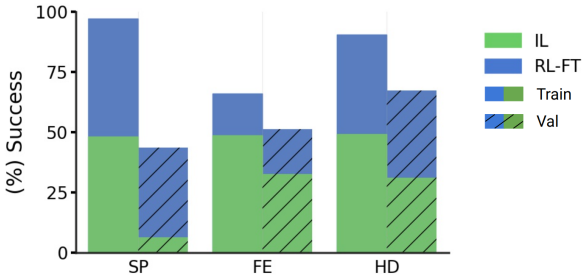

Figure 3. IL and RL performance for shortest paths (SP), frontier exploration (FE), and human demonstrations (HD) with equal IL training success on HM3D TRAIN (left) and VAL (right).

two factors that could be leading to this ordering after RL-finetuning – 1) inconsistency in performance at initialization (*i.e.* IL on HD is already better than IL on FE), and 2) amenability of each of these initializations to RL-finetuning (*i.e.* is RL-finetuning from HD init better than FE init?).

We are interested in answering (2), and so we control for (1) by selecting IL-pretrained policy weights across SP, FE, and HD that have equal performance on a subset of TRAIN $= \sim 48.0\%$ success. This essentially amounts to selecting IL-pretraining checkpoints for FE and HD from earlier in training as $\sim 48.0\%$ success is the maximum for SP.

Fig. 3 shows the results after IL and RL-finetuning on a subset of the HM3D TRAIN and on HM3D VAL. First, note that at IL-pretraining TRAIN success rates are equal ($= \sim 48.0\%$), while on VAL FE is slightly better than HD followed by SP. We find that after RL-finetuning, the policy trained on HD still leads to higher VAL success (66.1%) compared to FE (51.3%) and SP (43.6%). Notice that RL-finetuning from SP leads to high TRAIN success, but low VAL success, indicating significant overfitting. FE has smaller TRAIN-VAL gap after RL-finetuning but both are worse than HD, indicating underfitting. These results show that learning to imitate human demonstrations equips the agent with navigation strategies that enable better RL-finetuning generalization compared to imitating other kinds of demonstrations, even when controlled for the same IL-pretraining accuracy.

Results on SP-favoring and FE-favoring episodes. To further emphasize that imitating human demonstrations is key to good generalization at OBJECTNAV, we created two

Training demonstrations	IL Success (↑)	RL-FT Success (↑)
1) SP	**5.2**%	34.8%
2) HD	0.0%	**57.2**%
3) FE	**26.3**%	43.0%
4) HD	0.0%	**57.2**%

Table 2. Results on SP-favoring and FE-Favoring splits.

subsplits from the HM3D VAL split that are adversarial to HD performance – SP-favoring and FE-favoring. The SP-favoring VAL split consists of episodes where IL on SP achieved a higher performance compared to IL on HD, *i.e.* we select episodes where IL on SP succeeded but IL on HD did not or both IL on SP and IL on HD failed. Similarly, we also create an FE-favoring VAL split using the same sampling strategy biased towards IL on FE. Next, we report the performance of RL-finetuned from IL on SP, FE, and HD on these two evaluation splits in Table 2. On both SP-favoring and FE-favoring, IL on HD is at 0% success (by design), but after RL-finetuning, is able to significantly outperform RL-finetuning from the respective IL on SP and FE policies.

2.3. Scaling laws of IL and RL

In this section, we investigate how IL-pretraining and RL-finetuning performance scales with no. of IL demonstrations.

Human demonstrations. We create HD subsplits ranging in size from $2k$ to $77k$ episodes, and IL-pretrain policies with the same set of hyperparameters on each split. Then, for each, we RL-finetune from the best-performing checkpoint. The resulting IL and RL success on HM3D VAL *vs.* no. of HD episodes is plotted in Fig. 1. Similar to [1], we see promising scaling behavior with more IL demonstrations. Interestingly, as we increase the size of of the IL pretraining dataset and get to high IL accuracies, the improvements from RL-finetuning decrease. *E.g.* at $20k$ IL demonstrations, the IL→RL improvement is 10.1% success, while at $77k$ IL demonstrations, the improvement is 6.3%. Furthermore, with $35k$ IL-pretraining demonstrations, the RL-finetuned success is only 4% worse than RL-finetuning from $77k$ IL demonstrations (66.4% *vs.* 70.4%). Both of these suggest that by effectively leveraging the trade-off curve between the size of the IL-pretraining dataset *vs.* performance gains after RL-finetuning, it may be possible to achieve close to state-of-the-art results without large investments in demonstrations.

How well does FE Scale? In Sec. 2.1, we showed that IL on human demonstrations outperforms IL on both shortest paths and frontier exploration demonstrations, when controlled for the same amount of training experience. In contrast to human demonstrations however, collecting shortest paths and frontier exploration demonstrations is cheaper, which makes scaling these demonstration datasets easier. Since IL performance on shortest paths is significantly worse even with 3x more demonstrations compared to FE and HD ($240k$ SP *vs.* $70k$ FE and $77k$ HD demos, Sec. 2.1), we focus on scaling FE demonstrations. Fig. 4 plots performance on HM3D VAL against FE dataset size. We created splits ranging in size from $10k$ to $150k$. Increasing the dataset size doesn't consistently improve performance and saturates after $70k$ demonstrations, suggesting that simply generating more FE demonstrations is unlikely to help. We hypothesize that the saturation in performance is because these demonstrations do not capture task-specific exploration.

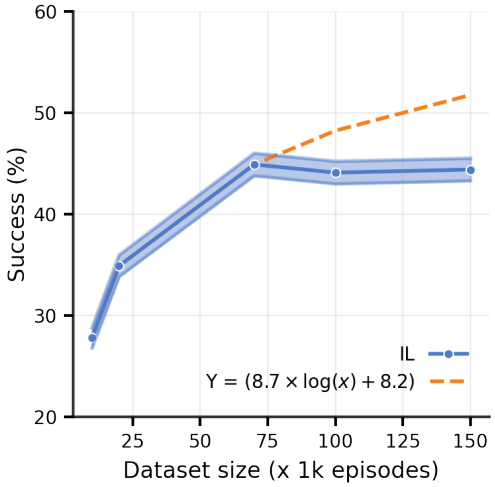

Figure 4. Success on ObjectNav HM3D VAL split vs. no. of frontier exploration demonstrations for training.

3. Conclusion

To conclude, we propose PIRLNav, an approach to combine imitation (IL) and reinforcement learning (RL) for OBJECTNAV, wherein we pretrain a policy with IL on $77k$ human demonstrations and then finetune it with RL, leading to state-of-the-art results on OBJECTNAV (65% success, 5% improvement over previous best). Next, using this IL+RL training recipe, we present a thorough empirical study of the impact of different demonstration datasets used for IL-pretraining on downstream RL-finetuning performance. We show that IL / IL+RL on human demonstrations outperforms IL / IL+RL on shortest paths and frontier exploration trajectories, even when we control for same IL success on TRAIN. We also show that as we scale the pretraining dataset size for IL and get to higher IL success rates, the improvements from RL-finetuning start to diminish. Finally, we characterize the failure modes of our best OBJECTNAV agent, and find that the largest sources of error are 1) dataset annotation noise (missing goal annotations, incorrect navigation meshes), and inability of the agent to 2) navigate across floors, and 3) recognize the correct goal object.

References

[1] R. Ramrakhya, E. Undersander, D. Batra, and A. Das, "Habitat-web: Learning embodied object-search strategies from human demonstrations at scale," in *CVPR*, 2022. 1, 2, 3, 4, 8, 10, 11, 13

[2] T. Winograd, "Understanding natural language," *Cognitive Psychology*, 1972. 1

[3] L. Smith and M. Gasser, "The development of embodied cognition: six lessons from babies.," *Artificial life*, vol. 11, no. 1-2, 2005. 1

[4] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. Czarnecki, M. Jaderberg, D. Teplyashin, *et al.*, "Grounded language learning in a simulated 3D world," *arXiv preprint arXiv:1706.06551*, 2017. 1

[5] F. Hill, K. M. Hermann, P. Blunsom, and S. Clark, "Understanding grounded language learning agents," *arXiv preprint arXiv:1710.09867*, 2017. 1

[6] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-attention architectures for task-oriented language grounding," in *AAAI*, 2018. 1

[7] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018. 1

[8] U. Jain, L. Weihs, E. Kolve, M. Rastegari, S. Lazebnik, A. Farhadi, A. G. Schwing, and A. Kembhavi, "Two body problem: Collaborative visual task completion," in *CVPR*, 2019. 1

[9] A. Das, *Building agents that can see, talk, and act*. PhD thesis, Georgia Institute of Technology, 2020. 1

[10] J. Abramson, A. Ahuja, I. Barr, A. Brussee, F. Carnevale, M. Cassin, R. Chhaparia, S. Clark, B. Damoc, A. Dudzik, *et al.*, "Imitating interactive intelligence," *arXiv preprint arXiv:2012.05672*, 2020. 1

[11] L. Weihs, A. Kembhavi, K. Ehsani, S. M. Pratt, W. Han, A. Herrasti, E. Kolve, D. Schwenk, R. Mottaghi, and A. Farhadi, "Learning generalizable visual representations via interactive gameplay," in *ICLR*, 2021. 1

[12] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," *arXiv preprint arXiv:2210.06407*, 2022. 1

[13] P. Anderson, A. X. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018. 1

[14] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "ObjectNav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020. 1, 8

[15] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D Data in Indoor Environments," in *3DV*, 2017. MatterPort3D dataset license: `http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf`. 2, 8

[16] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva, *et al.*, "Habitat-matterport 3d semantics dataset," *arXiv preprint arXiv:2210.05633*, 2022. 2, 8, 10, 14

[17] O. Maksymets, V. Cartillier, A. Gokaslan, E. Wijmans, W. Galuba, S. Lee, and D. Batra, "THDA: Treasure Hunt Data Augmentation for Semantic Navigation," in *ICCV*, 2021. 2, 9, 11

[18] S. Schaal, "Learning from demonstration," in *NIPS*, 1996. 2, 11

[19] A. Das, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Neural Modular Control for Embodied Question Answering," in *CoRL*, 2018. 2

[20] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *RSS*, 2018. 2, 11, 12

[21] B. Baker, I. Akkaya, P. Zhokhov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune, "Video pretraining (vpt): Learning to act by watching unlabeled online videos," *arXiv preprint arXiv:2206.11795*, 2022. 2, 10, 11, 13, 14

[22] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman, "Relay policy learning: Solving long horizon tasks via imitation and reinforcement learning," in *CoRL*, 2019. 2, 11

[23] I. Uchendu, T. Xiao, Y. Lu, B. Zhu, M. Yan, J. Simon, M. Bennice, C. Fu, C. Ma, J. Jiao, S. Levine, and K. Hausman, "Jump-start reinforcement learning," *arXiv preprint arXiv:2204.02372*, 2022. 2

[24] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *NeurIPS*, 2020. 2, 3, 10, 11

[25] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation'*, 1997. 2

[26] L. Weihs, U. Jain, I.-J. Liu, J. Salvador, S. Lazebnik, A. Kembhavi, and A. Schwing, "Bridging the imitation gap by adaptive insubordination," in *NeurIPS*, 2021. the first two authors contributed equally. 3

[27] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied Question Answering in Photorealistic Environments with Point Cloud Perception," in *CVPR*, 2019. 8

[28] K. Yadav, R. Ramrakhya, A. Majumdar, V.-P. Berges, S. Kuhar, D. Batra, A. Baevski, and O. Maksymets, "Offline visual representation learning for embodied navigation," 2022. 8, 11

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016. 8

[30] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *ICCV*, 2021. 8

[31] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021. 8

[32] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," *arXiv preprint arXiv:2107.09645*, 2021. 8

[33] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *ICLR*, 2020. 8

[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. 8, 12, 14

[35] S. K. Ramakrishnan, D. Jayaraman, and K. Grauman, "An exploration of embodied visual exploration," *arXiv preprint arXiv:2001.02192*, 2020. 10, 11, 14

[36] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," in *NeurIPS*, 2022. 10

[37] H. Team, "Habitat challenge, 2022." https://aihabitat.org/challenge/2022, 2020. 10

[38] J. Wasserman, K. Yadav, G. Chowdhary, A. Gupta, and U. Jain, "Last-mile embodied visual navigation," in *CoRL*, 2022. 11

[39] A. Mousavian, A.toshev, M. Fiser, J. Kosecka, A. Wahid, and J. Davidson, "Visual representations for semantic target driven navigation," in *ICRA*, 2019. 11

[40] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary Tasks and Exploration Enable ObjectNav," in *ICCV*, 2021. 11

[41] Y. Liang, B. Chen, and S. Song, "Sscnav: Confidence-aware semantic scene completion for visual semantic navigation," 2020. 11

[42] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning," in *CVPR*, 2022. 11

[43] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," in *ICLR*, 2019. 11

[44] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008. 11

[45] J. Kober and J. Peters, "Policy search for motor primitives in robotics," in *NeurIPS*, 2008. 11

[46] A. Nair, M. Dalal, A. Gupta, and S. Levine, "Accelerating online reinforcement learning with offline datasets," *arXiv preprint arXiv:2006.09359*, 2020. 11, 12

[47] Y. Lu, K. Hausman, Y. Chebotar, M. Yan, E. Jang, A. Herzog, T. Xiao, A. Irpan, M. Khansari, D. Kalashnikov, and S. Levine, "Aw-opt: Learning robotic skills with imitation andreinforcement at scale," in *CoRL*, 2021. 11, 12

[48] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *CoRL*, 2018. 11, 12

[49] X. B. Peng, A. Kumar, G. Zhang, and S. Levine, "Advantage-weighted regression: Simple and scalable off-policy reinforcement learning," *arXiv preprint arXiv:1910.00177*, 2019. 11, 12

[50] Q. Wang, J. Xiong, L. Han, p. sun, H. Liu, and T. Zhang, "Exponentially weighted imitation learning for batched historical data," in *NeurIPS*, 2018. 11, 12

[51] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, "Recurrent experience replay in distributed reinforcement learning," in *ICLR*, 2019. 11, 13

[52] K. Cobbe, J. Hilton, O. Klimov, and J. Schulman, "Phasic policy gradient," *arXiv preprint arXiv:2009.04416*, 2020. 11

[53] V. Kumar and E. Todorov, "Mujoco haptix: A virtual reality system for hand manipulation," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015. 12

[54] M. Hausknecht and P. Stone, "Deep Recurrent Q-Learning for Partially Observable MDPs," in *AAAI*, 2015. 12, 13

[55] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013. 12

[56] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *ICLR*, 2016. 13