Feudal Networks for Visual Navigation

Faith Johnson* Bryan Bo Cao† Ashwin Ashok‡ Shubham Jain† Kristin Dana*

Abstract

We introduce a novel no-RL, no-graph, no-odometry approach to visual navigation using feudal learning. This architecture employs a hierarchy of agents that each see a different aspect of the task and operate at different spatial and temporal scales. We develop two unique modules in this framework: (1) a ***memory proxy map*** *learned in a self-supervised manner that is used to record prior observations, and (2) a* ***waypoint network*** *that outputs intermediate subgoals by learning to imitate human waypoint selection during local navigation. This waypoint network is pre-trained using a dataset [1] of teleoperation sequences made publicly available in our prior work. The resulting feudal navigation network achieves SOTA performance on the image goal navigation task.*

Introduction Visual navigation is motivated by the idea in psychology that humans navigate with cognitive maps and graphs that preserve relative distances between landmarks [2–5] without ever building detailed 3D maps of their environment. In vision and robotics, these ideas have translated to the construction of topological graphs [6–10] and metric maps [11, 12] based primarily on visual observations [13–16]. Moreover, visual navigation methods seek new environment representations that are rich with semantic information [17–20], easy to dynamically update [21–23], and can be constructed faster and more compactly than full 3D metric maps [24–27]. NRNS [9] goes a step further by removing the reliance on simulators and reinforcement learning to train functional visual navigation models.

Our approach uses no simulator and no RL, but goes one step further by using no graphs and no odometry, resulting in a lightweight, easy-to-train visual navigation framework. We take inspiration from feudal learning [28–34], which identifies *workers* and *managers* and allows for multiple levels of hierarchy (ie. mid-level and high-level managers) that each observe different aspects of the task *and* operate at different temporal or spatial scales [35–37]. For navigation in unseen environments, this division of labor is ideal to make the overall task more manageable [38–40]. Our three tiered feudal navigation agent (FeudalNav) shown in Figure 1 achieves SOTA performance in image-goal navigation tasks in previously unseen Habitat [41] environments.

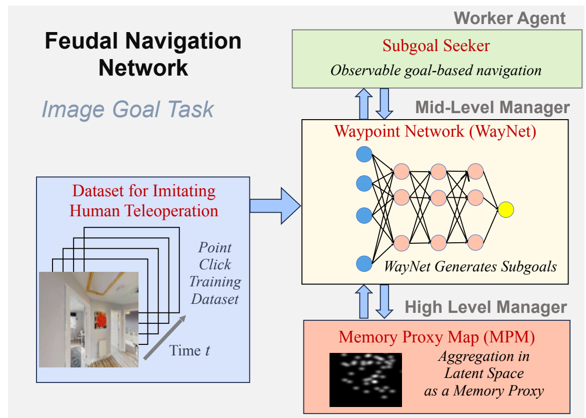

Figure 1. FeudalNav provides a no-graph, no-odometry, and no-RL visual navigation agent for the image-goal task on previously unseen environments. The hierarchy consists of: (1) a high-level manager with a memory proxy map (MPM) that frames memory as a latent space learning problem, (2) a mid-level manager waypoint network (WayNet) mimicking human teleoperation to guide worker agent exploration, and (3) a low-level worker choosing actions in the environment based on the previous layers' supervision.

Methods Key to our approach is representing traversed environments with a learned latent map that acts as a memory proxy during navigation. We contrastively learn a latent space that preserves the approximate distance between images to build an aggregate **memory proxy map (MPM)**. We learn this self-supervised latent space using a modified implementation of SMoG [46] that combines instance level contrastive learning and clustering methods. We add further modifications to model training in order to conduct navigation-aware, self-supervised contrastive learning on our Landmark-Aware Visual Navigation (LAVN) Dataset [1], which contains human waypoint-guided teleoperation trajectories in multiple virtual and real world environments. Instead of using typical constrastive learning data augmentation methods, we rely on the variations introduced through multiple camera views to learn robust image representations. During training, we build clusters for all trajectories where observations are grouped based on Superglue [47] robust keypoint matching. Then, we randomly sample positive pairs from each cluster to train the network.

As the agent navigates in novel environments, the high-level manager sequentially places observation images in this

* = Rutgers University, † = Stony Brook Univeristy, ‡ = Georgia State University
Corresponding Author: faith.johnson@rutgers.edu

Path Type	Model	Easy		Medium		Hard		Average	
		Succ↑	SPL↑	Succ↑	SPL↑	Succ↑	SPL↑	Succ↑	SPL↑
Straight	DDPPO (10M steps) * [42]	10.50	6.70	18.10	16.17	11.79	10.85	13.46	11.24
	DDPPO (extra data + 50M steps) * [42]	36.30	34.93	35.70	33.98	5.94	6.33	25.98	25.08
	DDPPO (extra data+100M steps) * [42]	43.20	38.54	36.40	34.89	7.44	7.20	29.01	26.88
	BC w/ ResNet + Metric Map [9]	24.80	23.94	11.50	11.28	1.36	1.26	12.55	12.16
	BC w/ ResNet + GRU [9]	34.90	33.43	17.60	17.05	6.08	5.93	19.53	18.80
	NRNS w/ noise [9]	64.10	55.43	47.90	39.54	25.19	18.09	45.73	37.69
	NRNS w/out noise [9]	68.00	61.62	49.10	44.56	23.82	18.28	46.97	41.49
	NRNS + SLING [43]	**85.3**	74.4	66.8	49.3	41.1	28.8	64.4	50.8
	OVRL + SLING [43]	71.2	54.1	60.3	44.4	43.0	29.1	58.2	42.5
	FeudalNav (Ours)	82.60	**74.95**	**71.00**	**57.40**	**49.01**	**34.20**	**67.54**	**55.52**
Curved	DDPPO (10M steps) * [42]	7.90	3.27	9.50	7.11	5.50	4.72	7.63	5.03
	DDPPO (extra data + 50M steps)* [42]	18.10	15.42	16.30	14.46	2.60	2.23	12.33	10.70
	DDPPO (extra data+100M steps)* [42]	22.20	16.51	20.70	18.52	4.20	3.71	15.70	12.91
	BC w/ ResNet + Metric Map [9]	3.10	2.53	0.80	0.71	0.20	0.16	1.37	1.13
	BC w/ ResNet + GRU [9]	3.60	2.86	1.10	0.91	0.50	0.36	1.73	1.38
	NRNS w/ noise [9]	27.30	10.55	23.10	10.35	10.50	5.61	20.30	8.84
	NRNS w/out noise [9]	35.50	18.38	23.90	12.08	12.50	6.84	23.97	12.43
	ZSEL* [20]	41.0	28.2	27.3	18.6	9.3	6.0	25.9	17.6
	OVRL* (53 GPU days) [44]	53.60	31.70	47.60	30.20	35.60	21.90	45.60	28.00
	NRNS + SLING [43]	58.6	16.1	47.6	16.8	24.9	10.1	43.7	14.3
	OVRL + SLING [43]	68.4	47.0	57.7	39.8	40.2	**25.5**	55.4	37.4
	FeudalNav (Ours)	**72.50**	**51.26**	**64.40**	**40.73**	**43.70**	25.32	**60.2**	**39.11**

Table 1. Quantitative comparison of our method (FeudalNav and Stacked FeudalNav) against baselines and SOTA on the image goal task following the evaluation protocol from NRNS [9] in previously unseen Gibson environments [45]. Bold = best performing.

latent space to dynamically build a memory proxy map of previously visited locations. We project SMoG features (128 dim) to a 2D latent space using a simple MLP that acts as an isomap imitator network by preserving the relative distance between image features. To update the MPM, we add a gaussian kernel to the corresponding 2D location in the map for each observation, thus creating a density map with values corresponding to the amount of exploration that has occurred in each location. The high-level manager polls the MPM's density to determine when a region is well-explored and movement away from the current region is desired.

The mid-level manager mimics human navigation policies by predicting a point in the environment to move towards. The intuition is that the human point-click navigation decisions in [1] are learnable and generalize to new environments with zero-shot transfer. We finetune Resnet-18 [48] to predict the pixel coordinate directing the navigation agent's motion in the environment from the combined input of the RGBD observation and the MPM. Navigation begins with Waynet predicting a waypoint for exploration. Concurrently, keypoint matches between the current observation and a goal image are computed by Superglue. If the confidence of this keypoint match is high, the average of the matched keypoints is used in the navigation pipeline instead of the waypoint prediction. In this manner the agent mimics human navigation in novel environments while checking if the goal location has been found.

The low-level worker agent chooses which actions to execute in the environment from the following action space: "turn left 15 degrees", "turn right 15 degrees", and "move forward 0.25 meters (m)". Although an RL agent is typically used for this type of task, we find a classifier works well to enable effective navigation. We train this classifier to learn a mapping between depth map and waypoint input and the corresponding human-chosen action from the LAVN dataset [1]. The agent chooses to stop navigation when the confidence threshold for matching goal image features to the current observation is high and either the agent's depth measurement indicates it is sufficiently close to the goal location or the area of the matched keypoints is relatively large with respect to the total image size.

Results We test the performance of FeudalNav using the procedure outlined in NRNS [9] on the image-goal task in previously unseen environments. All observation image are 480×640 pixels with 120° field of view. Each agent trajectory is evaluated on success rate (whether or not the agent reaches the goal) and SPL (success rate weighted by inverse path length). We compare FeudalNav's performance against a variety of SOTA methods in Table 1 and show improved performance to RL [42], behavior cloning [9], graph-based [9], last mile [43], zero-shot [20], and self-supervised [44] SOTA.

Acknowledgements

This work was supported by the National Science Foundation (NSF) under grant NSF NRT-FW-HTF: Socially Cognizant Robotics for a Technology Enhanced Society (SOCRATES) No. 2021628 and grant nos. CNS-2055520, CNS-1901355, CNS-1901133.

References

[1] Faith Johnson, Bryan Bo Cao, Kristin Dana, Shubham Jain, and Ashwin Ashok. A landmark-aware visual navigation dataset. *arXiv preprint arXiv:2402.14281*, 2024. 1, 2

[2] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948. 1

[3] Elizabeth R Chrastil and William H Warren. From cognitive maps to cognitive graphs. *PloS one*, 9(11):e112544, 2014.

[4] Michael Peer, Iva K Brunec, Nora S Newcombe, and Russell A Epstein. Structuring knowledge with cognitive maps and cognitive graphs. *Trends in cognitive sciences*, 25(1): 37–54, 2021.

[5] Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017. 1

[6] Kevin Chen, Juan Pablo De Vicente, Gabriel Sepulveda, Fei Xia, Alvaro Soto, Marynel Vázquez, and Silvio Savarese. A behavioral approach to visual navigation with graph localization networks. *arXiv preprint arXiv:1903.00445*, 2019. 1

[7] Dhruv Shah, Arjun Bhorkar, Hrish Leen, Ilya Kostrikov, Nick Rhinehart, and Sergey Levine. Offline reinforcement learning for visual navigation. *arXiv preprint arXiv:2212.08244*, 2022.

[8] Yuhang He, Irving Fang, Yiming Li, Rushi Bhavesh Shah, and Chen Feng. Metric-Free Exploration for Topological Mapping by Task and Motion Imitation in Feature Space. *arXiv preprint arXiv:2303.09192*, 2023.

[9] Meera Hahn, Devendra Singh Chaplot, Shubham Tulsiani, Mustafa Mukadam, James M Rehg, and Abhinav Gupta. No rl, no simulation: Learning to navigate without navigating. *Advances in Neural Information Processing Systems*, 34:26661–26673, 2021. 1, 2

[10] Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021. 1

[11] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning To Explore Using Active Neural SLAM. In *International Conference on Learning Representations*, 2019. 1

[12] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12875–12884, 2020. 1

[13] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive Mapping and Planning for Visual Navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[14] Alessandro Devo, Giacomo Mezzetti, Gabriele Costante, Mario L Fravolini, and Paolo Valigi. Towards generalization in target-driven visual navigation by using deep reinforcement learning. *IEEE Transactions on Robotics*, 36(5): 1546–1561, 2020.

[15] Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Ving: Learning open-world navigation with visual goals. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13215–13222. IEEE, 2021.

[16] Zachary Seymour, Kowshik Thopalli, Niluthpol Mithun, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Maast: Map attention with semantic transformers for efficient visual navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13223–13230. IEEE, 2021. 1

[17] Nuri Kim, Obin Kwon, Hwiyeon Yoo, Yunho Choi, Jeongho Park, and Songhwai Oh. Topological Semantic Graph Memory for Image-Goal Navigation. In *Conference on Robot Learning*, pages 393–402. PMLR, 2023. 1

[18] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. *arXiv preprint arXiv:2106.15648*, 2021.

[19] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.

[20] Ziad Al-Halah, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17031–17041, 2022. 1, 2

[21] Sacha Morin, Miguel Saavedra-Ruiz, and Liam Paull. One-4-All: Neural Potential Fields for Embodied Navigation. *arXiv preprint arXiv:2303.04011*, 2023. 1

[22] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022.

[23] Joao F Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018. 1

[24] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018. 1

[25] Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin,

Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31, 2018.

[26] Kevin Chen, Juan Pablo de Vicente, Gabriel Sepulveda, Fei Xia, Alvaro Soto, Marynel Vazquez, and Silvio Savarese. A Behavioral Approach to Visual Navigation with Graph Localization Networks. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, June 2019. doi: 10.15607/RSS.2019.XV.010.

[27] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991, 2023. 1

[28] Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. *Advances in neural information processing systems*, 5, 1992. 1

[29] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pages 3540–3549. PMLR, 2017.

[30] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.

[31] Ashish Kumar, Saurabh Gupta, David Fouhey, Sergey Levine, and Jitendra Malik. Visual memory for robust path following. *Advances in neural information processing systems*, 31, 2018.

[32] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 538–547, 2019.

[33] Edward Beeching, Jilles Dibangoye, Olivier Simonin, and Christian Wolf. EgoMap: Projective mapping and structured egocentric memory for Deep RL. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 525–540. Springer, 2020.

[34] Lina Mezghan, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3316–3323. IEEE, 2022. 1

[35] Alexander Vezhnevets, Yuhuai Wu, Maria Eckstein, Rémi Leblond, and Joel Z Leibo. Options as responses: Grounding behavioural hierarchies in multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 9733–9742. PMLR, 2020. 1

[36] Valerie Chen, Abhinav Gupta, and Kenneth Marino. Ask your humans: Using human instructions to improve generalization in reinforcement learning. *arXiv preprint arXiv:2011.00517*, 2020.

[37] Chengshu Li, Fei Xia, Roberto Martin-Martin, and Silvio Savarese. Hrl4in: Hierarchical reinforcement learning for interactive navigation with mobile manipulators. In *Conference on Robot Learning*, pages 603–616. PMLR, 2020. 1

[38] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning*, pages 1430–1440. PMLR, 2021. 1

[39] Chengguang Xu, Christopher Amato, and Lawson LS Wong. Hierarchical robot navigation in novel environments using rough 2-d maps. *arXiv preprint arXiv:2106.03665*, 2021.

[40] Jan Wöhlke, Felix Schmitt, and Herke van Hoof. Hierarchies of planning and reinforcement learning for robot navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10682–10688. IEEE, 2021. 1

[41] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1

[42] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 2

[43] Justin Wasserman, Karmesh Yadav, Girish Chowdhary, Abhinav Gupta, and Unnat Jain. Last-mile embodied visual navigation. In *Conference on Robot Learning*, pages 666–678. PMLR, 2023. 2

[44] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. *arXiv preprint arXiv:2204.13226*, 2022. 2

[45] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

[46] Bo Pang, Yifan Zhang, Yaoyi Li, Jia Cai, and Cewu Lu. Unsupervised visual representation learning by synchronous momentum grouping. In *European Conference on Computer Vision*, pages 265–282. Springer, 2022. 1

[47] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2