Mid-Level Vision at Habitat Challenge

Alexander Sax*[1] Jeffrey O. Zhang*[1] Bradley Emi[2] Amir Zamir[1,2]
Silvio Savarese[2] Leonidas Guibas[2] Jitendra Malik[1]

[1] University of California, Berkeley [2] Stanford University

http://perceptual.actor/

Abstract

Does knowing that the world is 3D help in delivering a package? More generally, how much does having ***visual priors about the world*** *assist in learning to perform* ***downstream motor tasks****? We study this question by integrating a generic perceptual skill set (e.g. a distance estimator, an edge detector, etc.) into a reinforcement learning framework—see Fig. 1. This skill set (hereafter* ***mid-level vision****) provides the policy with a more processed state of the world compared to raw images.*

We found that using mid-level vision confers significant advantages over training end-to-end without visual priors (i.e. tabula rasa*) in navigation-oriented tasks. The resulting policies generalize in situations where the* tabula rasa *approach fails, and the policies train with far fewer samples. However, realizing these gains requires careful selection of the mid-level visual skills. We refined this insight into an efficient* ***max-coverage feature set*** *that can be adopted in lieu of raw images. We submitted an policy trained with only mid-level vision to the 2019 CVPR Habitat Challenge, where it won the RGB track. The full paper, interactive visualizations, and an installable package implementing visual priors, are available at* http://perceptual.actor/.

1. Introduction

The renaissance of deep reinforcement learning (RL) started with the Atari DQN paper in which Mnih et al. [5] demonstrated an RL agent that learned to play video games directly from pixels. While the direct-from-pixels approach can learn arbitrary policies in an elegant, end-to-end fashion, there are two phenomena endemic to the paradigm: **I.** learning requires massive amounts of data (large sample complexity), and **II.** the resulting policies exhibit difficulties reproducing across environments with even modest visual differences (difficulty with generalization).

These phenomena are characteristic of a type of learning that is *too generic* [1]—in that it does not make use of *valid assumptions* that could boost generalization. Examples of such assumptions include that the world is spatially 3D or that certain groupings ("objects") behave together as a single entity. These are *facts* about the world and are generally true. Incorporating them as priors could provide an advantage over the assumption-free style of learning.

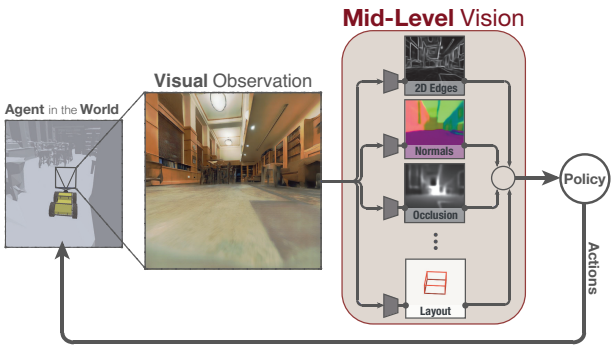

Figure 1. **Mid-level vision in an end-to-end framework for learning active robotic tasks.** Not incorporating such mid-level vision (i.e. bypassing the red box) is equivalent to learning directly from raw pixels. We report significant advantages in *sample efficiency* and *generalization* when using mid-level vision.

Including proper visual priors alleviates these two phenomena, improving both *generalization* and *sample efficiency*. We chose to implement visual priors by updating each policy's initial state from raw pixels to some mid-level features, using a neural network trained for some (isolated) computer vision task.

In the full paper, we showed that the best choice of prior depends on the downstream task, necessitating that we use a defensive *set* of features when we do not know what the downstream task will be *a priori*. Instead of hand-picking the features, we found this set computationally, presenting a **max-coverage feature set**. Experiments in Habitat [4] confirm that this performs roughly as well as the best-performing single feature.

Our full paper [6] contains extensive supporting evi-

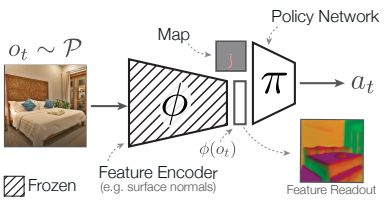

Figure 2. **Mid-level vision setup.** The learned features from fixed encoder networks are used as the state for training policies in RL. The setup here shows an agent using *surface normal* features. See readouts from each network as well as how the resulting polices perform, at our website

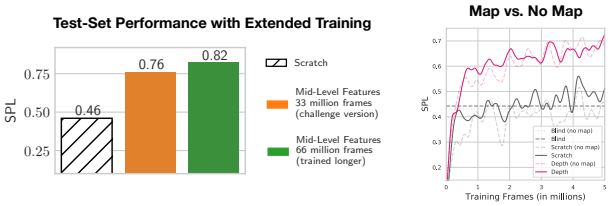

Figure 3. **Local planning in Habitat.** Agents with mid-level vision significantly outperform *scratch*. At the challenge deadline our agent had not yet finished training, and after continued training our agent achieved 0.06 higher SPL. *Right:* The location history provided by the map improved performance, but this gain was dwarfed by the improvement from using mid-level vision.

dence in the Gibson [11] and VizDoom [3] environments. Without any additional environment-specific tuning, we replicated those results in the Habitat [4] environment. This is a technical report that details the Habitat Challenge submission. For the full study, with both qualitative and quantitative interactive visualizations, an overview video, code, pretrained models, dockers containing our experiments, and an installable package implementing mid-level visual priors, please see our website.

2. Methodology and Experimental Setup

The Habitat Challenge consists of a *local planning* task where the agent must navigate to some target that is specified by a coordinate relative to the agent. The agent receives an RGB visual observation and a map of previously visited locations. This section details the high-level components of our submitted agent.

Mid-level visual priors: We implemented mid-level vision by taking the intermediate activations from a set of *encoder-decoder* neural networks that were each trained, offline, for a specific mid-level visual task–see Fig. 2. We froze each encoder's weights and used the encoder (ϕ) to transform each observed image o_t into a summary statistic $\phi(o_t)$ that we fed to the agent. During training, only the agent policy was updated.

We used the networks of [10] trained on a dataset of 4 million static images in of indoor scenes [10]. Each network encoder consists of a ResNet-50 [2] without a global average-pooling layer. See our website for a full list of features and sample videos of networks evaluated in our environments.

Non-visual components: This details the high-level non-visual components of our agent. For an exhaustive description of the mid-level setup as well as hyperparameter settings and the code, please see our website.

Evaluation metric: Performance is measured using the *Success Weighted by Path Length* (SPL) [4] metric where 0 is the worst and 1 is the best possible.

RL algorithm: In all experiments we use the common Proximal Policy Optimization (PPO) [8] algorithm with Generalized Advantage Estimation [7]. We decorrelate our batches using experience replay and off-policy variant of PPO, similar to Actor-Critic with Experience Replay (ACER) [9]. See [6] for the full formulation.

No tuning: We do no hyperparameter tuning, instead using the parameters from the *local planning* task in Gibson from [6], which were optimized for *tabula rasa* learning.

Single-GPU: We trained our mid-level feature-based agents using a single GPU, which we used for both training and environment rollouts.

Backout and jerk validation: Our original submission used a takeover policy that assumes control from the trained agent when the agent gets stuck or tries to move Left/Right multiple times in a row. This improved performance by 0.03 SPL, but we achieved larger gains by simply training longer after the challenge.

3. Results

Without any hyperparameter tuning and using only the features found useful in [6], the feature-based agent achieved an SPL of 0.82[1] By comparison, agents trained *tabula rasa* achieved an SPL of around 0.46 and these results are completely consistent with our findings in [6].

In addition to achieving higher generalization performance, feature-based agents were much more sample efficient than learning *tabula rasa* (Fig. 4). Feature-based agents reached the maximum reward achieved by *tabula rasa* with only 1/10th of the training data.

We note in Fig. 3 that the map improved final performance, but this change is *in addition* to the benefits of mid-level vision and almost all of the SPL improvement comes from the visual priors. In fact, the memoryless feature-based agent significantly outperforms the *scratch* agent, even when that agent has access to the map.

In terms of *which* features fared well in the challenge: they were not necessarily the ones currently considered essential (Fig. 4). Although *depth* performed well, others features such as *curvature* performed comparably or better.

Finally, we used the **Max-Coverage Feature Set** from the full paper, finding that it performed about the same as our best agent (*curvature*). These results are shown in Fig. 3. This, again, is consistent with our findings in [6].

[1] This is 0.06 SPL higher than our submission, which managed to win the RGB track of the challenge despite not being fully trained.

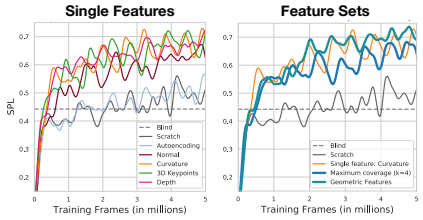

Figure 4. **Comparison of various types of mid-level vision in Habitat.** *Left:* Agents using *depth* features outperform *scratch*, but other representation perform comparably or better. However, not all are useful and *autoencoder* features perform no better than *scratch*. *Right:* Our task-agnostic *Max-Coverage Feature Set* performs roughly as well as the best features. However, task-specific feature sets might have even better performance (geometric set).

4. Discussion

Including mid-level visual priors closed two-thirds of the gap between direct-from-pixels RL and the best-possible agent. This would suggest that one of the primary difficulties with RL-from-pixels is using the pixels themselves. Upgrading the visual input from pixels to something more processed, such as encodings, has clear benefits. However, the exactly what those encodings should be and how to choose them for each task is not yet clear. The generic *Max-Coverage Min-Distance* set of mid-level features performed well in the Habitat Challenge–about as well as the the best single feature.

More in-depth analysis and a discussion of limitations is given in the full study: http://perceptual.actor/.

Acknowledgements We gratefully acknowledge the support of ONR MURI (N00014-14-1-0671) and NSF (IIS-1763268), NVIDIA NGC beta, and TRI. Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, Jan 1992. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[3] M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jaskowski. Vizdoom: A doom-based AI research platform for visual reinforcement learning. *CoRR*, abs/1605.02097, 2016. 2

[4] O. Y. Z. E. W. B. J. J. S. J. L. V. K. J. M. D. P. Manolis Savva*, Abhishek Kadian* and D. Batra. Habitat: A platform for embodied ai research. *arXiv preprint arXiv:1904.01201*, 2019. 1, 2

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015. 1

[6] A. Sax, B. Emi, A. R. Zamir, L. J. Guibas, S. Savarese, and J. Malik. Mid-level visual representations improve generalization and sample efficiency for learning active tasks. *CoRR*, abs/1812.11971, 2018. 1, 2

[7] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *CoRR*, abs/1506.02438, 2015. 2

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. 2

[9] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas. Sample efficient actor-critic with experience replay. *CoRR*, abs/1611.01224, 2016. 2

[10] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2

[11] A. R. Zamir, F. Xia, J. He, A. Sax, J. Malik, and S. Savarese. Gibson Env: Real-world perception for embodied agents. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2