Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Data Science Lab Fall 25

ETH AI CENTER

Develop meta-RL policy for varying morphologies using proxy and task training

David Scherrer [1], Constantin Pinkl[1], Fatemeh Zargarbashi [2], Arnout Devos [3]
[1]ETH Zurich, D-INFK [2]ETH Computational Robotics Lab [3]ETH AI Center, ETH Zurich

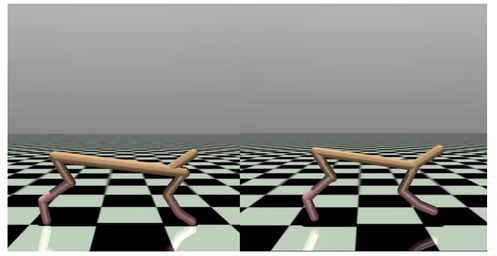

Figure 1: Different Cheetah robots with various body variations (Brockman et al., 2016)

1. Problem Statement

Develop a meta-RL policy that generalizes to unseen HalfCheetah morphologies at test time by training across diverse embodiments and leveraging a proxy task.

2. Related work

Our work builds on Zargarbashi et al. (2024), which uses a GRU-based meta-RL architecture to encode morphology in its hidden state. We extend this with a proxy-task pretraining phase, where the agent first solves a sinusoidal height-modulation task that encourages morphology-specific structure.
Bohlinger et al. (2024) handle varying morphologies via explicit encoding of physical attributes and an attention-based feed-forward policy. In contrast, our method learns morphology implicitly through recurrent interaction history.

3. Data

We rely on simulated interaction data obtained from MuJoCo-based OpenAI Gym environments. Using the Gymnasium HalfCheetah blueprint, we alter its body proportions to generate diverse embodiments that challenge the agent to generalize across morphologies.

4. Methods

We train a recurrent meta-RL policy that first solves a proxy task and then the main locomotion task. During training, the GRU processes continuous interaction data across both phases, allowing its hidden state to capture morphology-dependent dynamics.

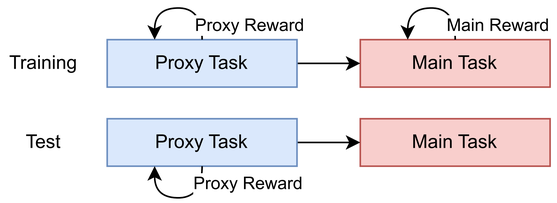

Figure 2: During training, the agent solves the proxy task before the main task, maintaining the GRU hidden state across phases. At test time, the same sequence is executed, and the agent continues to receive proxy rewards while optimizing the main task.

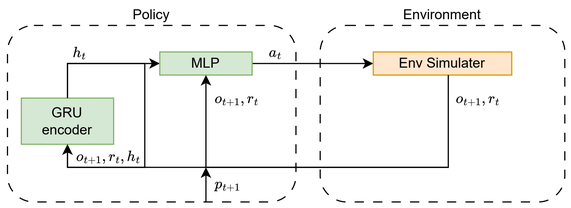

Figure 3: The GRU integrates observations, rewards, and the current training phase into a hidden state that summarizes the robot's dynamics. This hidden state, together with the current observation, reward and phase is passed through an MLP to generate actions. The environment returns the next observation and reward, forming the recurrent interaction cycle.

4.1 Proxy Task – Sinusoidal Height Tracking

The agent learns to follow a sinusoidal target height trajectory by raising and lowering its torso. This task encourages morphology-aware control, joint coordination, and stabilization. Rewards combine tracking accuracy and a small velocity penalty:

$$r_{\text{proxy}}(t) = w_1 \exp\left(-\frac{(h(t) - h_{\text{target}}(t))^2}{\sigma_h}\right) + w_2 \exp\left(-\frac{v(t)^2}{\sigma_v}\right)$$

$$h_{\text{target}}(t) = h_{\text{base}} + A \sin\left(\frac{2\pi t}{T}\right).$$

4.2 Main Task – Locomotion Forward Task

The downstream objective is fast and stable forward locomotion. The reward promotes forward velocity and upright posture:

$$r_{\text{main}}(t) = w_3 \exp\left(\frac{v_x(t)}{\sigma_f}\right) + w_4 \exp\left(\frac{\max(0,\ z_{\text{torso}}(t) \cdot z_{\text{world}} - m)}{(1-m)\sigma_u}\right).$$

5. Results

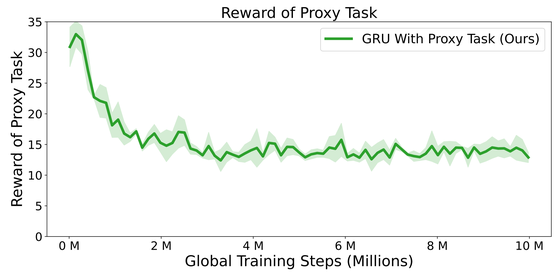

Figure 4: Proxy task reward over training, averaged across all evaluation scenarios. Only *GRU With Proxy Task* achieves non-zero proxy reward, which decreases as the agent shifts focus toward the main task. The experiment was conducted over 3 seeds.

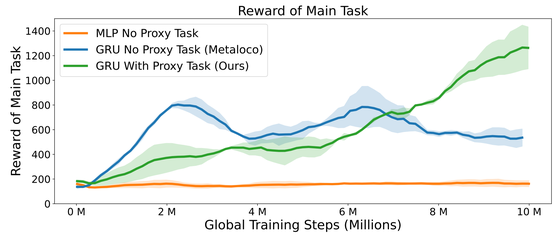

Figure 5: Main task reward during training, averaged across all evaluation scenarios. *GRU With Proxy Task* achieves the highest final performance, surpassing *GRU No Proxy Task*, which peaks early and then declines. *MLP No Proxy Task* shows slow and limited improvement, highlighting the advantage of recurrent policies and proxy-based training. The experiment was conducted over 3 seeds.

6. Conclusions

- **Proxy Task** *GRU WithProxy* shows strong initial proxy performance that degrades over time, indicating that the agent gradually stops exploiting the proxy signal.
- **Main Task** Leveraging the proxy ultimately enables the GRU agent to reach higher final performance than both the GRU without proxy and the MLP baseline.

7. Future Work

- Extend the approach to more complex quadruped robots.
- Explore additional proxy tasks to study their impact.

References

Nico Bohlinger, Grzegorz Czechmanowski, Maciej Krupka, Piotr Kicki, Krzysztof Walas, Jan Peters, and Davide Tateo. One policy to run them all: an end-to-end learning approach to multi-embodiment locomotion. *arXiv preprint arXiv:2409.06366*, 2024.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Fatemeh Zargarbashi, Fabrizio Di Giuro, Jin Cheng, Dongho Kang, Bhavya Sukhija, and Stelian Coros. Metaloco: Universal quadrupedal locomotion with meta-reinforcement learning and motion imitation, 2024. URL `https://arxiv.org/abs/2407.17502`.

Contributions

DS, CS: Methodology, Software, Visualization, Investigation, Writing - Original Draft
FZ: Conceptualization, Methodology, Supervision
AD: Supervision, Methodology, Project Administration

Code. Associated code is available open-source and under the MIT License, which permits free reuse, free modification, and free distribution, provided proper attribution is given to the authors, and no liability is assumed by the authors. Follow up work is not required to be open-source and is not required to have an MIT License. The code and its MIT license are publicly available here: `[https://github.com/djscherrer/robodogs`