Advanced Skills through Multiple Adversarial Motion Priors in Reinforcement Learning

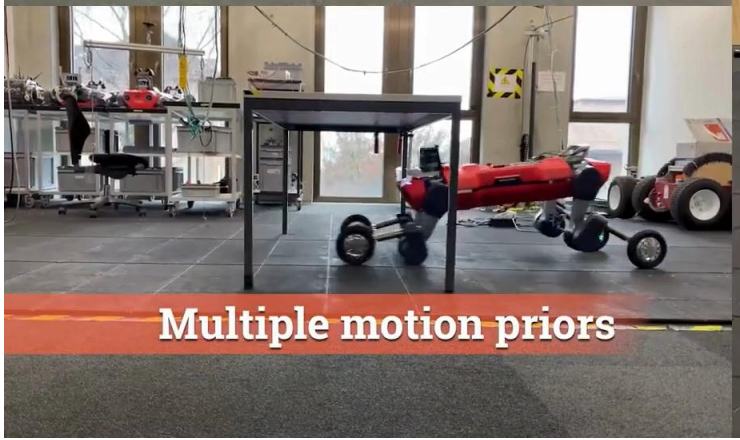
Paper by Eric Vollenweider, Marko Bjelonic, Victor Klemm, Nikita Rudin, Joonho Lee and Marco Hutter

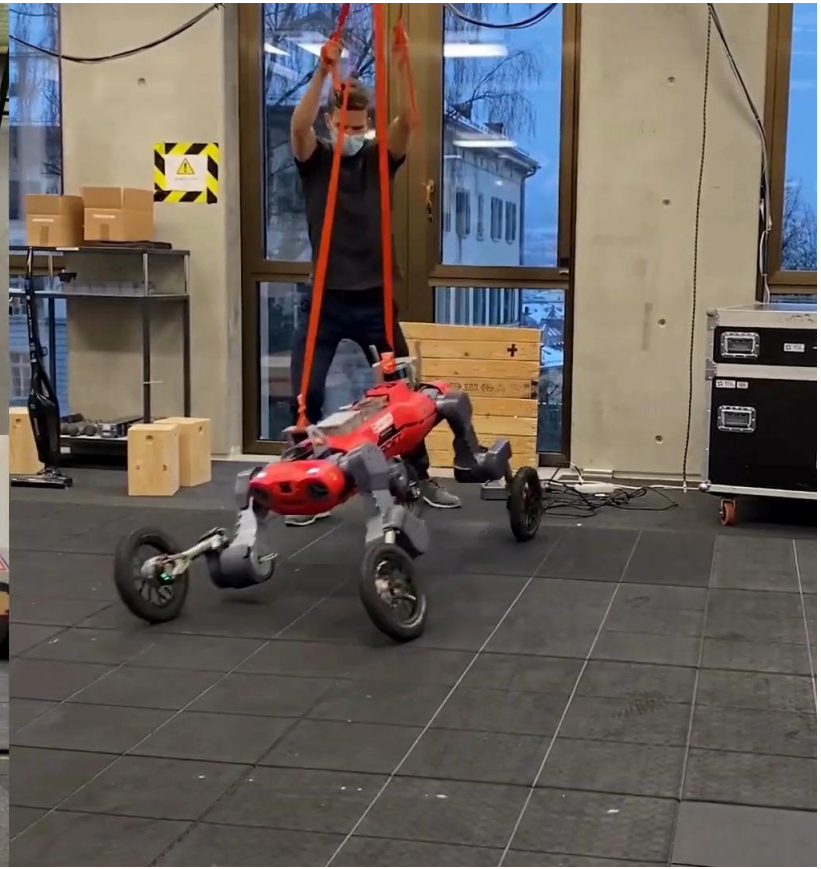
[1]

Matt Anner

Estelle Baumann

Johanne Pinel


Multiple motion priors



Idea/contribution of the article

Challenge

Tedious process of tuning the reward function with RL approaches

Idea

Imitation learning with Adversial Motion Priors (AMP) with multiple discretely switchables motions styles

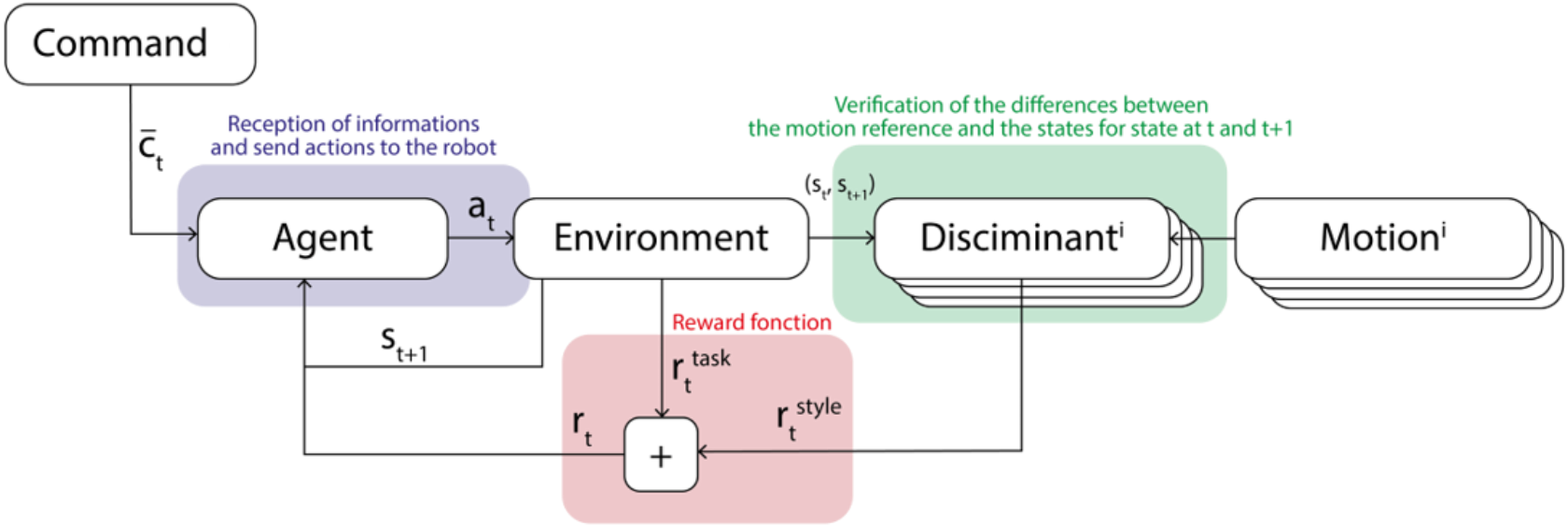
Task

Four-legged, ducking, and standing up locomotion

Robot

ANYmal on Wheels robots (16 DOF) and torques controlled

Multiple adversarial motion priors



- Goal : learn differents style of motion with one policy of renforcement

AMP and Multi-AMP

Reward Calculation

$r = r_{task} + r_{style}$

How to do it

What to do

motivates to extract the motion style from given data

i : one of the trained styles

The motion dataset of style i (= reference style)

M^i

Discriminator, a neural network that compares B to M

D^i

Style reward :
Rewards high when they look similar
Rewards low when they don't

Buffer which stores the robot's attempts to perform style i

B_π^i

Each of the trainable style is defined by a tuple :

$\{D^i, B_\pi^i, M^i\}$

Require: $M = \{M_i\}, |M| = n$ (n motion data-sets)

$\pi \leftarrow$ initialize policy
$V \leftarrow$ initialize Value function
$[\mathcal{B}] \leftarrow$ initialize n style replay buffers
$[D] \leftarrow$ initialize n discriminators
$\mathcal{R} \leftarrow$ initialize main replay buffers
while not done **do**
  for trajectory i = 1, ..., m **do**
    $\tau^i \leftarrow \{(c_t, c_s, s_t, a_t, r_t^G)_{t=0}^{T-1}, s_T, g\}$ roll-out with π
    $d \leftarrow$ style-index of τ^i (encoded in c_s)
    if M^d is not empty **then**
      for t = 0, ..., T-1 **do**
        $d_t \leftarrow D^d(\phi(s_t), \phi(s_{t+1}))$
        $r_t^{style} \leftarrow$ according to Eq. 2
        record r_t^{style} in τ^i
      end for
      store d_t in $\mathcal{B}^d$ and τ_i in $\mathcal{R}$
    end if
  end for
  for update step = 1, ..., $n_{updates}$ **do**
    for d = 0, ..., n **do**
      $b^{\mathcal{M}} \leftarrow$ sample batch of K transitions $\{s_j, s'_j\}_{j=1}^K$ from $\mathcal{M}^d$
      $b^{\pi} \leftarrow$ sample batch of K transitions $\{s_j, s'_j\}_{j=1}^K$ from $\mathcal{B}^d$
      update D^d according to Eq. 1
    end for
  end for
  update V and π (standard PPO step using $\mathcal{R}$)
end while

Require: $M = \{M_i\}, |M| = n$ (n motion data-sets)

$\pi \leftarrow$ initialize policy
$V \leftarrow$ initialize Value function
$[\mathcal{B}] \leftarrow$ initialize n style replay buffers
$[D] \leftarrow$ initialize n discriminators
$\mathcal{R} \leftarrow$ initialize main replay buffers
while not done **do**
  for trajectory $i = 1, ..., m$ **do**
    $\tau^i \leftarrow \{(c_t, c_s, s_t, a_t, r_t^G)_{t=0}^{T-1}, s_T, g\}$ roll-out with π
    $d \leftarrow$ style-index of τ^i (encoded in c_s)
    if M^d is not empty **then**
      for $t = 0, ...,$ T-1 **do**
        $d_t \leftarrow D^d(\phi(s_t), \phi(s_{t+1}))$
        $r_t^{style} \leftarrow$ according to Eq. 2
        record r_t^{style} in τ^i
      end for
      store d_t in $\mathcal{B}^d$ and τ_i in $\mathcal{R}$
    end if
  end for
  for update step $= 1, ..., n_{updates}$ **do**
    for d $= 0, ...,$ n **do**
      $b^{\mathcal{M}} \leftarrow$ sample batch of K transitions $\{s_j, s'_j\}_{j=1}^K$ from $\mathcal{M}^d$
      $b^{\pi} \leftarrow$ sample batch of K transitions $\{s_j, s'_j\}_{j=1}^K$ from $\mathcal{B}^d$
      update D^d according to Eq. 1
    end for
  end for
  update V and π (standard PPO step using $\mathcal{R}$)
end while

C_t : command, contains C_s
C_s : one-hot-encoded selector, its elements are zero everywhere except at the index of style i

Collect trajectories τ_i by running π in the environment + r_{task} computation

$$r_t^{task} = R(c_t, s_t, s_{t-1})$$

we have motion data

Compute r_{style} using the discriminator for this style

$$r_t^{style} = -log\left(1 - \frac{1}{1 + \exp^{-D^i([\phi(s_t), \phi(s_{t+1})])}}\right)$$

Discriminator update iteration :
To make sure the discriminator does not fall behind the updated policy

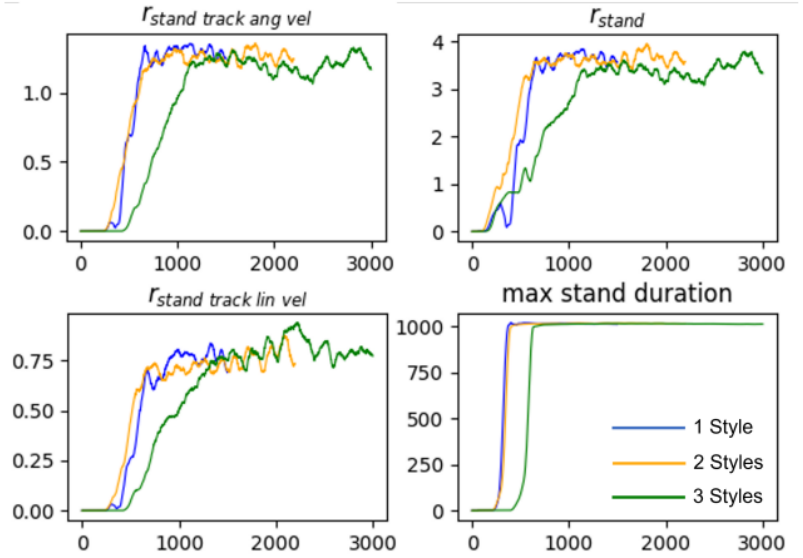
Experimentations

- Sim-to-real with Wheeled-Legged robot (ANYmal on Wheels) with 16 DOFs

- 3 tasks for the training environment :
  - Four-legged locomotion
  - Duck under a table
  - Standing up & Sitting down

Difference between simulations and real robot [2]

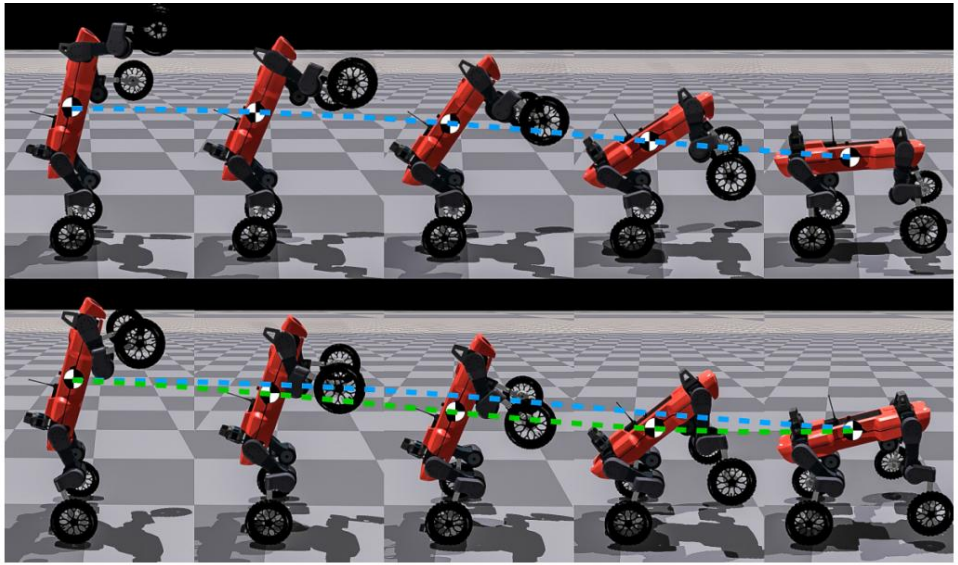
Results



Multi-AMP learning capability of the stand up task. Horizontal axis denotes the number of epochs and vertical axis represents the rewards after post-procesing for comparability [2]

Difference between simulations and real robot [2]

Switching between quadruped and humanoid configuration



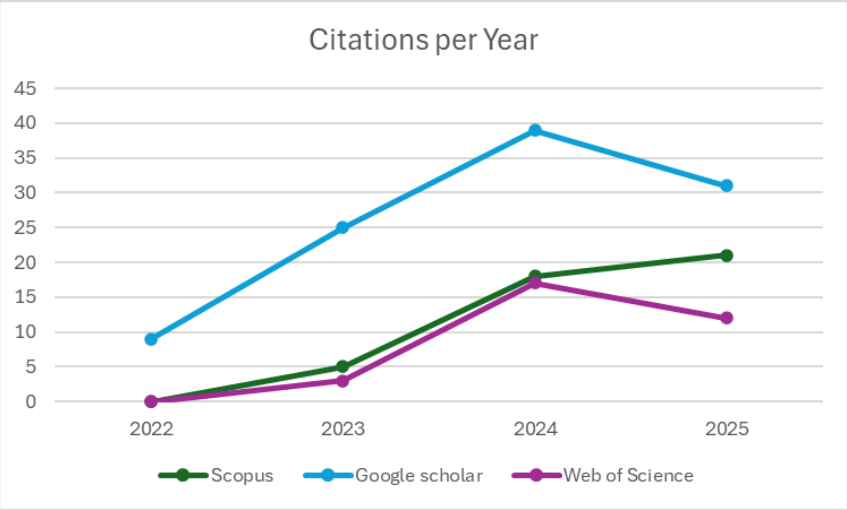
In blue the trajectory of the center of mass with the task rewards only, in green the trajectory of the center of mass with the reversed stand-up sequence. [2]

[3]

- Problem: Sit-down motion exceeds safety torques threshold
- Solution: Record stand-up motion, reverse the motion data and then train with Mutli-AMP

Influence of the paper

- No further works on Multi-AMP as such
- Works on extension on the idea of learning multiple skill as one (CAMP [4], SMSL [5])
- The article is often cited in synthesis article about
- First of its kind to do sim-to-real



Pros and Cons

Pros

- Training of multiple skill in one simulation
- No reward function tuning
- No differences for walking and ducking together in regards of alone

Cons

- More time required to generates motion priors
- Learning of the 3 styles together is a bit longer to archive
- Sit-down works in simulation but exceed safety values in real-life

References

- [1] *Interesting Engineering. (2023, Dec 6). Boston Dynamics' Atlas robot performing parkour*
- [2] Vollenweider, N., *et al.* (2023). *Advanced Skills through Multiple Adversarial Motion Priors in Reinforcement Learning*. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 29 – June 2, 2023, London, UK.
- [3] Extracts from YouTube. *(2023, June 15). How drones are revolutionizing live events [Video]. YouTube. https://www.youtube.com/watch?v=kEdr0ARq48A*
- [4] N. Huang, Z. Xie, and Q. Li, "Learning Multi-Skill Legged Locomotion Using Conditional Adversarial Motion Priors," Sept. 26, 2025, *arXiv*: arXiv:2509.21810. doi: 10.48550/arXiv.2509.21810.
- [5] J. Tu, P. Zhai, Y. Zhang, X. Wei, Z. Dong, and L. Zhang, "Seamless multi-skill learning: learning and transitioning non-similar skills in quadruped robots with limited data," *Front. Robot. AI*, vol. 12, Apr. 2025, doi: 10.3389/frobt.2025.1542692.