Figure 1: Video of real-world testing

Blind Bipedal Stair Traversal via Sim-to-Real Reinforcement Learning

Camille Coppieters De Gibson
Chiara Evangelisti
Advaith Sriram

Introduction

- **Keywords**: Bipedal Locomotion, Sim-to-Real Transfer, Long Short-Term Memory (LSTM), Proprioception
- **Research aim**: To make bipedal robots capable of navigating stairs using only proprioception, without relying on fragile external sensors like cameras
- **Sim-to-Real Transfer**: The training occurred entirely in simulation, randomizing stair features such as height, width, and slope, before transferring the policy to real-world testing

Figure 2: Cassie – Bipedal robot by Orgeon State University and Agility Robotics

First successful demonstration of a bipedal robot traversing real-world stairs using only proprioception and RL

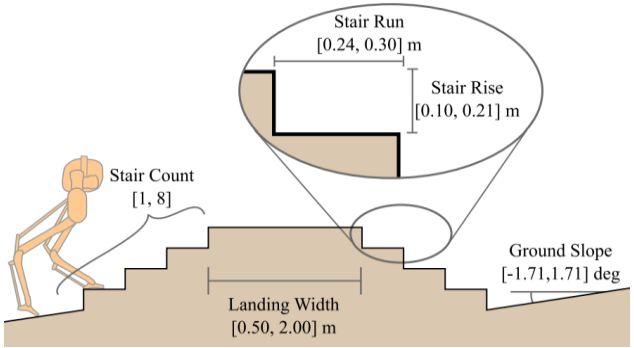


Figure 3: Randomization of stair features

Key aspects

- Biped robot, human-scale
- Actuators: electrical
- Control: torque in the joints
- Sensors: Blind robot has no exteroceptive sensors
- Design: learns through sim-to-real Reinforcement Learning
  - Built on reward-function
- Navigate Staircase with no specific gait

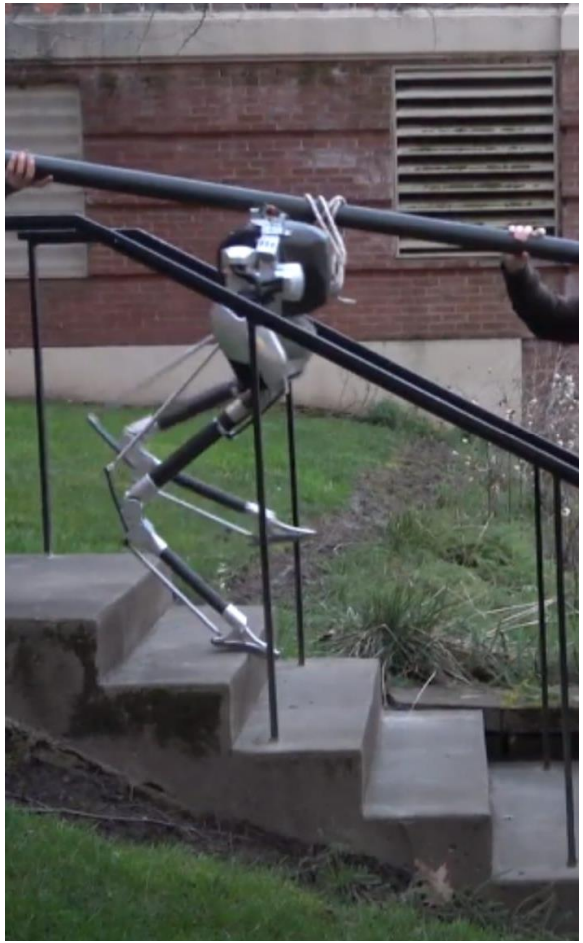


Figure 4: Real-world testing performed

Reinforcement learning formulation (1)

- **State space**:
  1) Robot's *physical state*: pelvis orientation and angular speed, joint positions and velocities
  2) *High level command inputs* from human operator with randomization
  3) *2 Cyclic clock inputs:* to manage gait phases

$$p = \begin{cases} \sin(2\pi(\phi_t + 0.0)) \\ \sin(2\pi(\phi_t + 0.5)) \end{cases}$$

- **Action space:**
  1) 10 *PD targets* for the joints
  2) *Clock delta:* to regulate stepping frequency of the gait

$$\phi_{t+1} = \text{fmod}(\phi_t + \delta_t, 1.0)$$

↳ Limited impact on performance

Reinforcement learning formulation (2)

- **Dynamics randomization**: several dynamics quantities are *randomized* at the beginning of each episode to overcome possible *modelling errors*

Parameter	Unit	Range
Joint damping	Nms/rad	$[0.5, 3.5] \times$ default values
Joint mass	kg	$[0.5, 1.7] \times$ default values
Ground Friction	–	$[0.5, 1.1]$
Joint Encoder Offset	rad	$[-0.05, 0.05]$
Execution Rate	Hz	$[37, 42]$

- **Policy representation and learning:**
  - I. LSTM recurrent neural network for the goal policy
  - II. Feedforward neural network for ablation experiment

Memory improves performance for Partially Observable environments

Proximal policy optimization (PPO), KL-threshold-termination variant, mirror loss term for symmetric gait

Reinforcement learning formulation (3)

- **Reward function:** $R(s,\phi) = 1 - \mathbb{E}[\rho(s,\phi)]$
  - I. Terms involving *expectation:* vary during gait cycle, *penalization* foot forces and foot velocities *at key intervals* → periodic foot lift
  - II. Terms to match a translational velocity and *orientation*
  - III. Terms to improve smoothness and energy efficiency and reduce pelvis shakiness
- **Terrain randomization:**
  - I. Stairs dimensions + noise
  - II. Ground friction
  - III. Stairs starting position

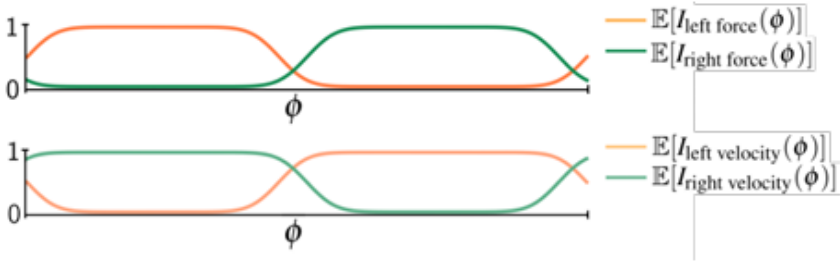


Figure 5: Cyclic coeffient for the reward: expectation of random indicator functions of the phase

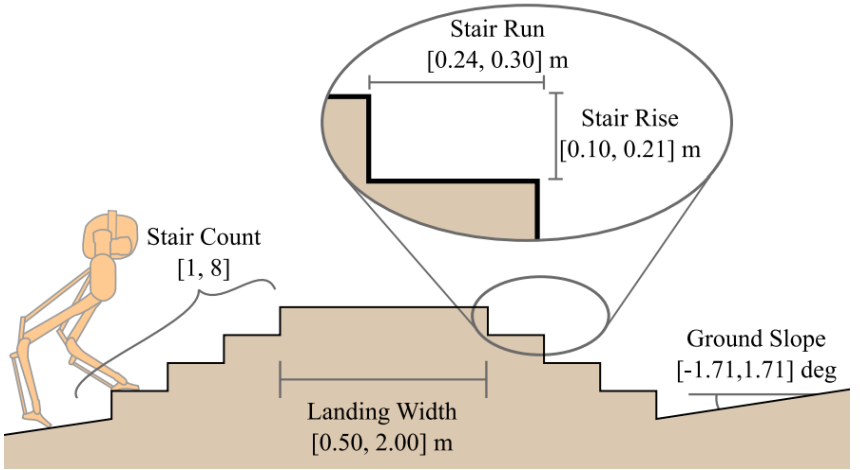


Figure 6: Randomization of stair features

Main results (1)

- **4 policies groups trained:**
  1. Stair LSTM
  2. Stair FF (feedforward)
  3. Flat ground LSTM
  4. Proximity LSTM → 1) + information nearby stairs

- **Simulation results**
  1. Probability of successfully ascending/descending stairs
  2. Energy efficiency

- **Behavior analysis:** 1st step up/down
  1. Swing foot motion
  2. Ground reaction forces

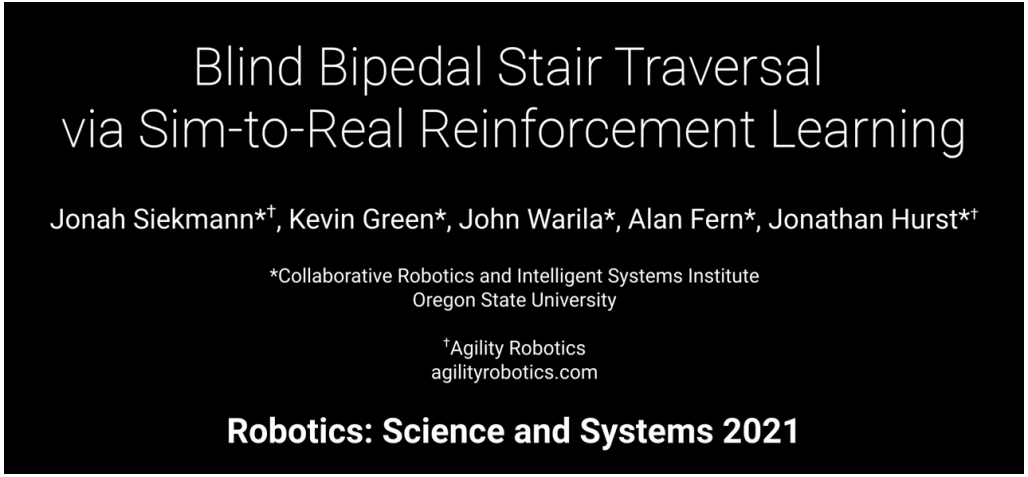


Figure 7: Video – Simulation training and real-world transfer

Main results (2): Ascending / descending probabilities

Success = reaching the top/ bottom without falling

- 150 trials,
- 5 steps,
- 17 cm tread,
- 30 cm depth

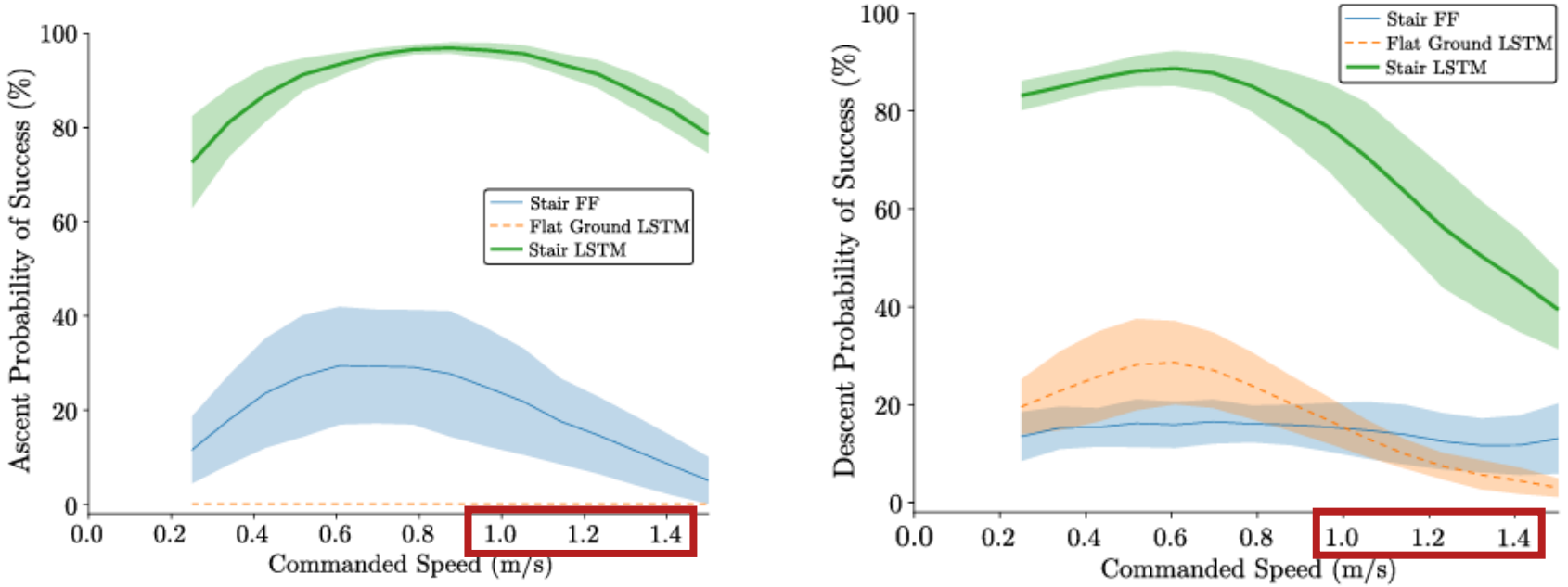


Figure 8: Comparison of success probabilities of step ascent and descent as a function of commanded speed between different policies

Main results (3): Energy efficiency comparison

Measure : Cost of transport ➡ $\mathrm{CoT} = \frac{E_\mathrm{m}}{Mgd}$

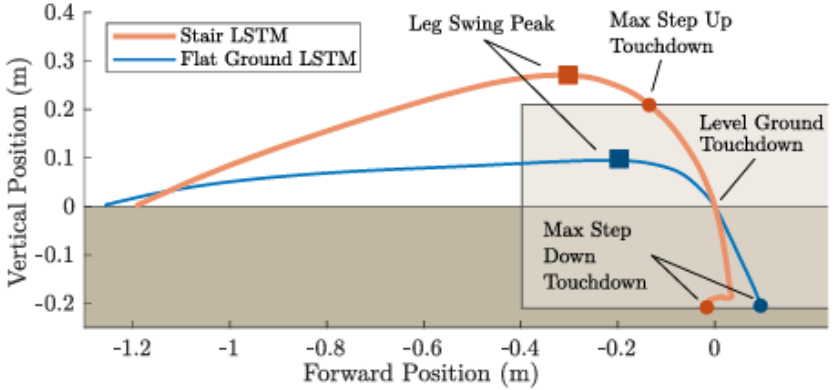
Cassie energy consumption : positive actuator work and resistive losses

$$E_\mathrm{m} = \int_0^T \left(\sum_i max(\tau_i \cdot \omega_i, 0) + \frac{\omega_i^{\max}}{P_i^{\max}} \tau_i^2 \right) dt$$

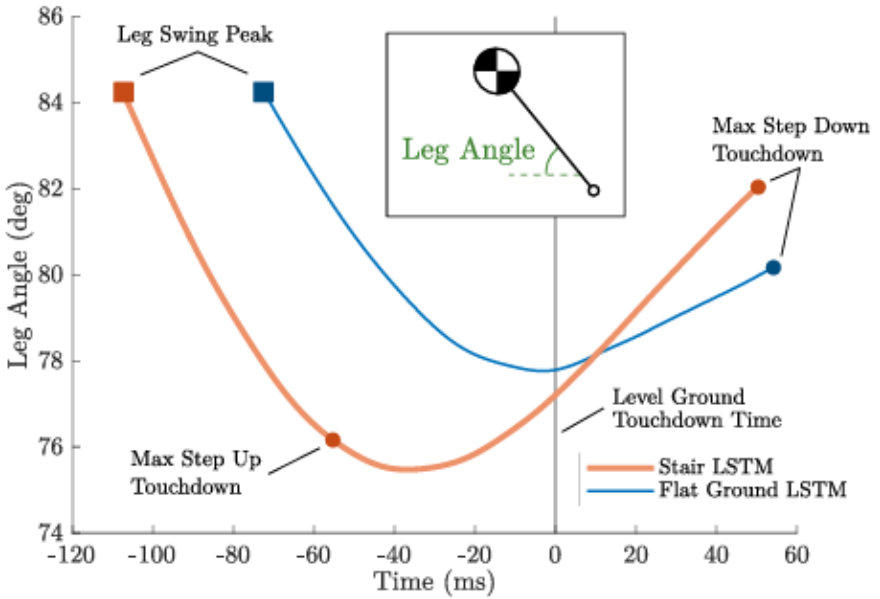
Testing: 1 m/s on flat ground

Policy Group	Mean CoT	Std. CoT
Proximity LSTM (stairs)	0.47	0.0086
Stair LSTM	0.46	0.0323
Proximity LSTM (flat)	0.39	0.0257
Flat Ground LSTM	0.38	0.0205

Main results (4) : Swing foot motion



(a) Swing foot paths for the stair trained policy and the flat ground policy overlaid on example step ups and step downs.



(b) The leg angle between the robot body and the swing foot as the foot descends toward touchdown.

Stair LSTM:
- higher step
- steeper path
- faster leg retraction rate

Emerging leg swing retraction even if not trained

Figure 9: Comparison of swing foot motion between different policies while locomoting at 1 m/s

Main results (5) : Transfer to hardware

- Robust, error-correcting , reliable behavior for stairs → missteps
- Robustness to unven terrains, logs and curbs → **not trained**
- Robustness to inclines and deformable terrains → wet grass field, small hill
- **80%** success for ascending, **100**% for descending

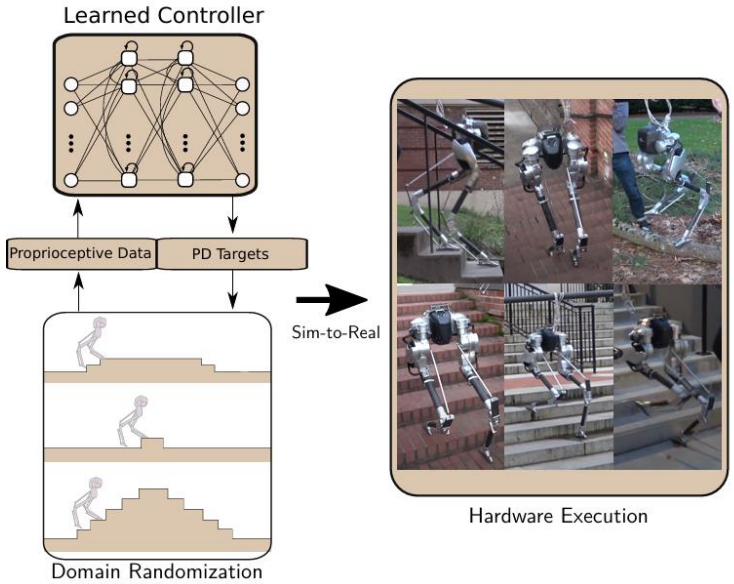


Figure 10: Training pipeline from simulation to hardware

Influence on Literature

Citations: 194
FWCI: 12.69

- Bipedal Locomotion:
  - Cassie cited as one of the few successful biped robots handling outdoor rough terrains, despite high hardware demands
  - Described stair-climbing as a fundamental skill for future bipedal research

- Randomization of Dynamics:
  - Positively referenced on numerous occasions for the use of Domain Randomization
  - Some research found better results with FF policies vs LSTM in narrow randomization ranges (Singh et al., 2023)

- Quadrupedal Locomotion:
  - Mentions of robustness, benefits of sim-to-real transfer and RL implementation
  - Criticism of use of only proprioceptive sensors and speed of robot (Takahiro et al., 2022) (Margolis et al., 2024)

Advantages and limitations

Advantages

- Highly robust, visionless walking controller
- Climbs a wide variety of real-world stairs
- Use existing reward functions
- Didn't require additional information to build a control policy either
- adapted well to other unknown terrains beyond stairs
- Random tests proved highly effective

Limitations

- Need of memory
- High energy requirement
- Higher failure rate on slatted stairs

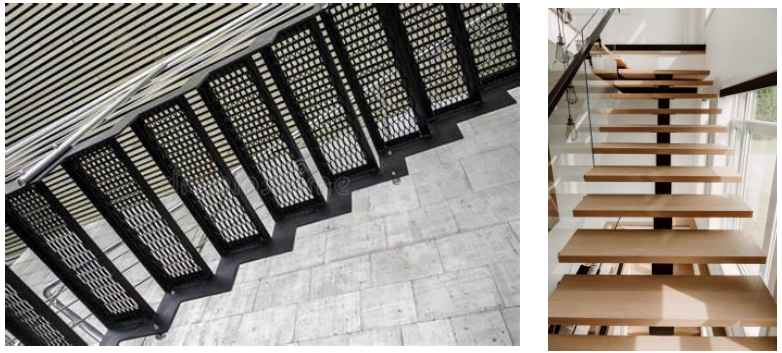


Figure 11: Slatted stairs

Possible exam questions

- **Which is the main innovative contribution of this work?**

The authors managed to obtain a highly robust walking controller capable of climbing a variety of stairs using only proprioceptive sensing (blind robot). This was achieved using an existing reward strategy for walking without adding any stair-specific rewards terms, but just adding randomized stairs to the training environment. A key feature to learn this behavior is a memory component in the policy representation.

- **How is the success rate for ascending/descending linked to the approach speed? (Slide 9)**

The policies fail more frequently at low approach speed as they might not have enough momentum to compensate and push the robot when the foot placement is not ideal. Similarly higher failure rates are observed at high speeds due to a high momentum which makes it harder to control precisely in case of small errors in foot placement or leg angles. Besides reduced contact time and increased forces are also possible causes for a less stable gait.

Additional slides

Details on reinforcement learning formulation

- **Randomization of control commands in the state space**:

Command	Probability of Change	Range
Forward Speed	1/300	[-0.3m/s, 1.5m/s]
Sideways Speed	1/300	[-0.3m/s, 0.3m/s
Turn Rate	1/300	[-90deg/s, 90deg/s]

At each timestep they are altered with a given probability

New command is sampled from uniform distribution

Advantage: give policies during training a good variety of speeds and approach angles to start with

- **Policy representation and learning**:

LSTM recurrent NN: 2 reccurrent hidden layers of dimension 128. Batches of episodes

Feedforward NN : two layers of dimension 300, with tanh activation. Batches of timesteps

In this method proposed by Yu et al.[Yu et al. 2018], the create a symmetry loss defined as follows:

$$L_{sym}(\theta) = \sum_{t=1}^{T} \|\pi_\theta(s_t) - M_a(\pi_\theta(M_s(s_t)))\|^2$$

and optimize this as an auxiliary loss in addition to the def loss:

$$\pi_\theta = \underset{\theta}{\operatorname{argmin}}\, L_{PPO}(\theta) + wL_{sym}(\theta),$$

Figure 11: Reference for mirror loss component in PPO

300 timesteps per episode, 7.5 s

Details on reinforcement learning formulation

- **Reward function**

Weight	Cost Component
0.140	$1 - \mathbb{E}[I_{\text{left force}}(\phi)] \cdot \exp(-.01\|F_l\|)$
0.140	$1 - \mathbb{E}[I_{\text{right force}}(\phi)] \cdot \exp(-.01\|F_r\|)$
0.140	$1 - \mathbb{E}[I_{\text{left velocity}}(\phi)] \cdot \exp(-\|v_l\|)$
0.140	$1 - \mathbb{E}[I_{\text{right velocity}}(\phi)] \cdot \exp(-\|v_r\|)$
0.140	$1 - \exp(-\varepsilon_o)$
0.140	$1 - \exp(-\lvert\dot{x}_{\text{desired}} - \dot{x}_{\text{actual}}\rvert)$
0.078	$1 - \exp(-\lvert\dot{y}_{\text{desired}} - \dot{y}_{\text{actual}}\rvert)$
0.028	$1 - \exp(-5 \cdot \|a_t - a_{t-1}\|)$
0.028	$1 - \exp(-0.05 \cdot \|\tau\|)$
0.028	$1 - \exp(-0.1(\|\text{pelvis}_{\text{rot}}\| + \|\text{pelvis}_{\text{acc}}\|))$

$$\varepsilon_o = 3(1 - \hat{q}^T q_{\text{body}})^2 + 10\left((1 - \hat{q}^T q_{\text{l}})^2 + (1 - \hat{q}^T q_{\text{r}})^2\right)$$

$I_i(\phi)$ Binary-valued random indicator function
1 → active interval
0 → inactive interval

Expectation used for **more stable learning**

Details on training settings

- 300 millions timesteps sampled from simulation using Mujoco
- 50’000 timesteps for replay buffer
- 64 episodes (LSTM) / 1024 (FF) batch size
- Replay buffer sampled up to 5 epochs
- Optimization termination if 0.2 maximum KL threshould reached
- Adam optimizer, 0.0005 learning rate for both actor and critic (learned separately)

Main results (6) : Ground reaction forces, 10 cm step

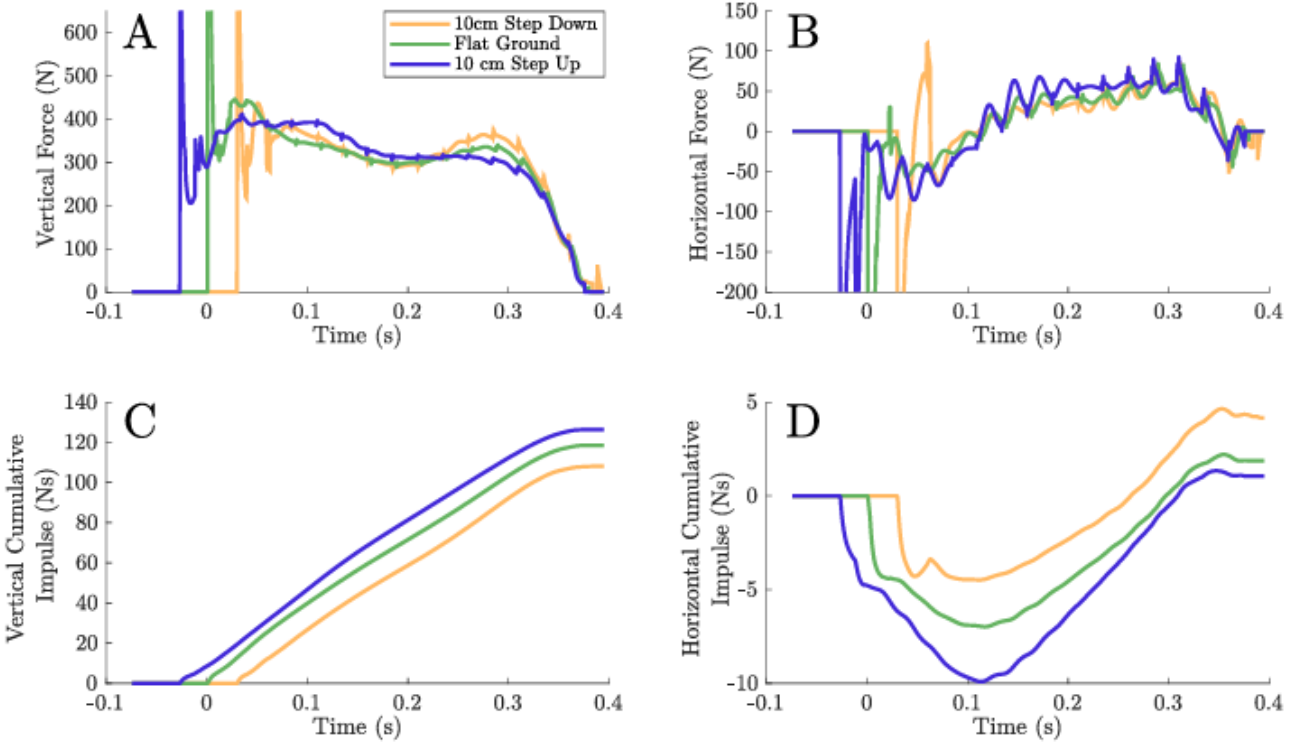


Figure 12: Comparison of ground reaction forces and cumulative impulses between different policies when going up or down a 10 cm step

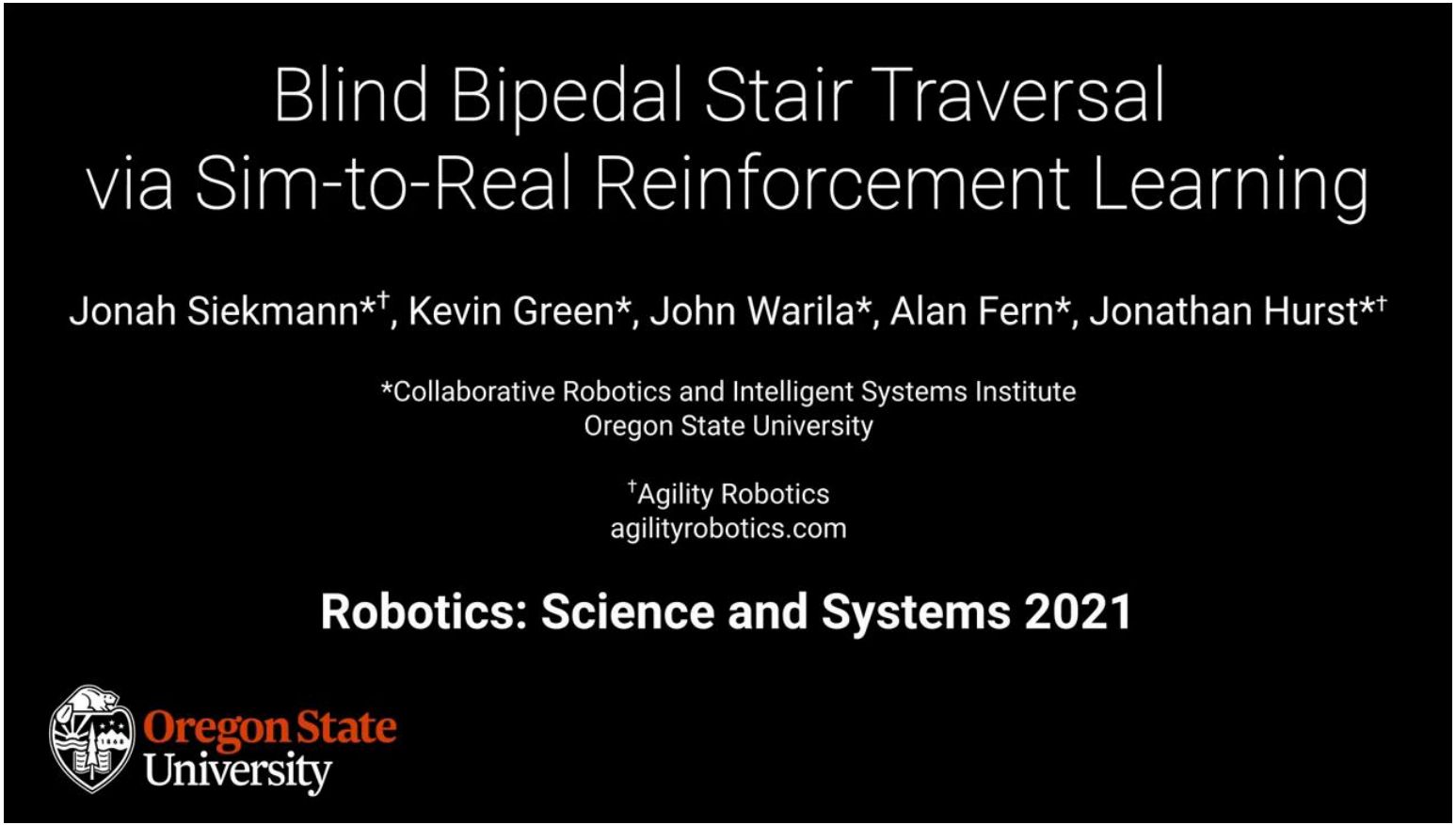
A. Maximum nominal leg force quite **constant** → *well adjusted policy* **Increased force** second bump for descent

B. Oscillations match policy evaluation **frequency** → Hp: policy controlling pelvis attitude

C. **Higher** vertical impulse for stepping up → cheaper to lift down

D. **Larger** horizontal impulse for stepping down → predicted by leg swing retraction

Video


Blind Bipedal Stair Traversal
via Sim-to-Real Reinforcement Learning
Jonah Siekmann*†, Kevin Green*, John Warila*, Alan Fern*, Jonathan Hurst*†
*Collaborative Robotics and Intelligent Systems Institute
Oregon State University
†Agility Robotics
agilityrobotics.com
Robotics: Science and Systems 2021
Oregon State University