


DeepGait: Planning and Control of Quadrupedal Gaits using Deep Reinforcement Learning

Journal Article**Author(s):**

Tsounis, Vassilios; Alge, Mitja; Lee, Joonho; Farshidian, Farbod; [Hutter, Marco](#) 

Publication date:

2020-04

Permanent link:

<https://doi.org/10.3929/ethz-b-000404175>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

Originally published in:

IEEE Robotics and Automation Letters 5(2), <https://doi.org/10.1109/LRA.2020.2979660>

DeepGait: Planning and Control of Quadrupedal Gaits using Deep Reinforcement Learning

Vassilios Tsounis*, Mitja Alge*, Joonho Lee, Farbod Farshidian, and Marco Hutter

Abstract—This paper addresses the problem of legged locomotion in non-flat terrain. As legged robots such as quadrupeds are to be deployed in terrains with geometries which are difficult to model and predict, the need arises to equip them with the capability to generalize well to unforeseen situations. In this work, we propose a novel technique for training neural-network policies for terrain-aware locomotion, which combines state-of-the-art methods for model-based motion planning and reinforcement learning. Our approach is centered on formulating Markov decision processes using the evaluation of dynamic feasibility criteria in place of physical simulation. We thus employ policy-gradient methods to independently train policies which respectively plan and execute foothold and base motions in 3D environments using both proprioceptive and exteroceptive measurements. We apply our method within a challenging suite of simulated terrain scenarios which contain features such as narrow bridges, gaps and stepping-stones, and train policies which succeed in locomoting effectively in all cases.

Index Terms—Legged Robots; Deep Learning in Robotics and Automation; Motion and Path Planning

I. INTRODUCTION

LEGGED locomotion in non-flat terrain, both structured and unstructured, poses a significant challenge in robotics. Operating autonomously in such environments requires addressing the problem of multi-contact motion planning and control. If a legged robot such as ANYmal [1] is to traverse complex environments autonomously, it must possess the capability to select footholds appropriate for the terrain, while also retaining balance at all times. This work deals specifically with the problem of planning and executing sequences of footholds for quadrupedal locomotion in rigid non-flat terrain using proprioceptive and exteroceptive sensing.

Dynamically walking on non-flat terrain necessitates the optimization of continuous state-input trajectories such as the motion of the base, as well as discrete decision variables such as which surface, and when, to make contact with. This has been addressed predominantly using model-based approaches, such as those employing deterministic optimization

Manuscript received: September, 10, 2019; Revised December, 17, 2019; Accepted January, 18, 2020.

This paper was recommended for publication by Editor Nikolaos G. Tsagarakis upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by Intel Labs, the Swiss National Science Foundation (SNSF) through project 166232, 188596, the National Centre of Competence in Research Robotics (NCCR Robotics), and the European Union's Horizon 2020 research and innovation program under grant agreement No.780883. Moreover, this work has been conducted as part of ANYmal Research, a community to advance legged robotics.

All authors are with the Robotic Systems Lab, ETH Zürich, Switzerland. tsounisv@ethz.ch

* These authors contributed equally.

Digital Object Identifier (DOI): see top of this page.

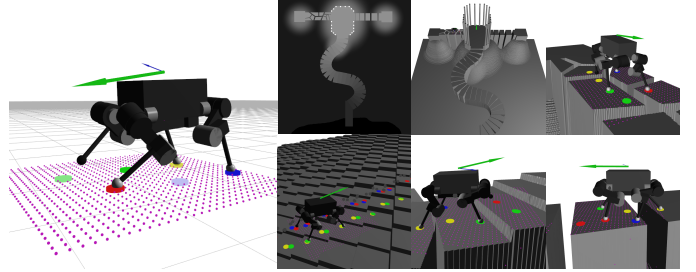


Fig. 1: The suite of terrains: the baseline Flat-World scenario (left), the Random-Stairs scenario (bottom center), and composite Temple-Ascent (right) scenario comprising a set of winding stairs and two derelict bridges containing stepping-stones and gaps of varying size.

techniques [2], [3], in conjunction with other heuristics [4], to plan motions for both the base and feet. Although some of the aforementioned approaches are able to solve such problems compromising both continuous and discrete variables [2], [3], these remain too computationally intensive to be executed online. Thus, only kinostatic approaches [4], [5] have managed to perform online foothold planning. Also, most methods typically employ some form of parameterization or qualification of the terrain [4], [6], [7] in order to simplify the search for viable footholds.

One of the primary challenges in multi-contact planning for multi-legged systems lies in dealing with the combinatorial problem due to the vast number of contact configurations admissible for the terrain. Typical solutions involve either assuming the gait pattern [4], [5] or employing sampling-based search techniques [7], [8]. There also exist works that have combined both optimization and sampling-based methods [5], [6], [9]. However, these typically resort to decoupling the selection of footholds from the optimization of base motions and thus remain kinostatic as they tend to neglect the dynamics of the system.

Some works have also incorporated machine-learning techniques for facilitating terrain perception [6], [10], [11]. Others, have employed Deep Reinforcement Learning (DRL) [12], [13] for realizing end-to-end terrain-aware locomotion. The use of the latter, however, still poses several challenges, namely: (1) how to eliminate undesirable yet retain beneficial emergent behavior, and (2) reduce overall sample complexity and train policies efficiently.

We propose a new method that combines state-of-the-art model-based and model-free methods to enable quadrupedal systems to traverse complex non-flat terrain. Our formulation consists of: (1) a terrain-aware planner that generates sequences of footholds and base motions that direct the robot towards a target heading, and (2) a foothold and base motion

controller which executes the aforementioned sequence while maintaining balance and dealing with disturbances. Both planner and controller are realized as stochastic policies parameterized using Neural-Network (NN) function approximation, which are optimized using state-of-the-art Deep Reinforcement Learning (DRL) algorithms.

Our contributions with this work are: (1) A novel method for training kinodynamic motion planners, which employs a Trajectory Optimization (TO) technique for determining so-called *transition feasibility* between discrete support phases using a coarse model of the robot. This removes the need for a planner to interact with both physics and a controller during training, allows the two policies to be trained independently, and leads to a significant reduction in overall sample complexity. (2) A simple formulation for realizing dynamic walking controllers that use target footholds as references and rely solely on proprioceptive sensing. This allows us to train controllers that can fully exploit the kinematics and dynamics of the robot in order to track arbitrary target footholds, irrespective of the planner used to generate them.

We evaluate the performance of our method across a set of challenging locomotion scenarios using a physics simulator and present results thereof. Our experiments demonstrate that the planner can generalize well across terrain types, and the controller succeeds in tracking reference footholds while always balancing the robot. Moreover, we illustrate the advantages of our method by comparing it with a state-of-the-art model-based approach [4].

II. PRELIMINARIES

A. Reinforcement Learning

We consider the problem of sequential decision making in which an agent interacts with an environment with the objective of maximizing cumulative reward. We model this problem as a discrete-time infinite Markov Decision Processes (MDP) with a discounted expected return objective. Such an MDP consists of set of states \mathcal{S} , a set of actions \mathcal{A} , a transition dynamics distribution, an initial state distribution, a scalar reward function $r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$, and a scalar discount factor $\gamma \in [0, 1)$. The agent selects actions according to a policy π with the objective of maximizing the expected return $\mathbb{E}[\sum_{k=t}^{\infty} \gamma^k r_{t+k}]$, where r_t is the scalar reward resulting from the state transition at time-step t . As we consider infinite MDPs in which \mathcal{S} and \mathcal{A} are infinite sets, we use parameterized stochastic policies $\pi_{\theta}(\mathbf{a}|\mathbf{o}_t)$, which are distributions over actions $\mathbf{a} \in \mathcal{A}$ conditioned on observations $\mathbf{o}_t \in \mathcal{O}$ given parameter vectors $\theta \in \mathbb{R}^n$.

B. Model of the System

The robot comprises an unactuated floating base and four articulated legs with actuated rotational joints. The state of the robot is specified as: $\mathbf{r}_B \in \mathbb{R}^3$ the absolute¹ position of the base, $\mathbf{R}_B \in SO(3)$ is the rotation matrix representing the

absolute attitude of the base, $\mathbf{v}_B, \boldsymbol{\omega}_B \in \mathbb{R}^3$ are the absolute linear and angular velocities of the base, $\mathbf{q}_j, \dot{\mathbf{q}}_j \in \mathbb{R}^{12}$ are the angular positions and velocities of the joints. The robot is controlled using joint torques $\boldsymbol{\tau}_j \in \mathbb{R}^{12}$. Moreover, we assume that we can extract robocentric measurements of *local* terrain elevation via the mapping $\mathbf{M}_R : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^{32 \times 32}$ with a resolution of 4 cm. In order to reason precisely about gaits and transitions between contact supports, we define a parameterization thereof that encompasses all necessary information. We thus parameterize a gait as a sequence of so-called *support phases*. Each phase is defined by the tuple

$$\Phi := \langle \mathbf{R}_B, \mathbf{r}_B, \mathbf{v}_B, \mathbf{r}_F, \mathbf{c}_F, t_E, t_S \rangle \in \Phi \quad (1)$$

where $\mathbf{c}_F \in \{0, 1\}^4$ is a vector indicating for each of the feet a closed, 1, or open, 0, contact w.r.t the terrain, $\mathbf{r}_F \in \mathbb{R}^{3 \times 4}$ are the stacked absolute positions of the feet, and $t_E, t_S \in \mathbb{R}$ are the phase timing variables. For every phase Φ_t , $t - t_E$ defines the time at which the switch to the current contact configuration occurred, while $t + t_S$ the switch to next. Fig. 2 (b) illustrates the aforementioned quantities.

III. METHODOLOGY

We propose a two-level hierarchy comprising a high-level Gait Planner (GP) and a low-level Gait Controller (GC) operating at different time-scales, inspired by [13]. The GP, evaluated at roughly 2Hz, serves as a local terrain-aware planner, and uses both *exteroceptive* and *proprioceptive* measurements to generate a finite sequence of support phases, i.e. a *phase plan*. The GC, evaluated at 100Hz, serves as a hybrid motion planner and controller, and uses *only proprioceptive* sensing in combination with the aforementioned phase plan to output joint position references. Finally, a joint-space PD controller (with zero target joint velocity) uses these joint position references to compute joint torques at 400Hz. The highest-level command to the system is provided as the desired walking direction in the form of the deviation of base attitude w.r.t the goal. Fig. 2(a) provides an overview of the system.

A. Gait Planning

The GP operates by sequentially querying the planning policy π_{θ_p} to generate the aforementioned phase plan. We thus formulate an MDP in order to train π_{θ_p} using DRL, and our objective is to ensure that the resulting policy learns to respect the kinodynamic properties and limits of the robot, as well as contact constraints, when proposing phase transitions. Moreover, we aim to avoid direct interaction with the physics of the system, and instead craft the transition dynamics of the MDP by employing a *transition feasibility* criterion realized as a Linear Program (LP) using the framework defined in [14]. Lastly, we avoid explicitly modeling or qualifying the terrain, as done in [4], [8], and instead directly use measurements of local terrain elevation. The resulting MDP, allows us to train π_{θ_p} to infer a distribution over phase transitions, which are not only feasible but also maximize locomotion performance (refer to the discussion in Sec. V).

¹We define a global inertial frame W for the world, and local body-fixed frames B for the base and F for the feet. Left sub-scripts denote the frame in which the vector is expressed, but omit it for absolute quantities.

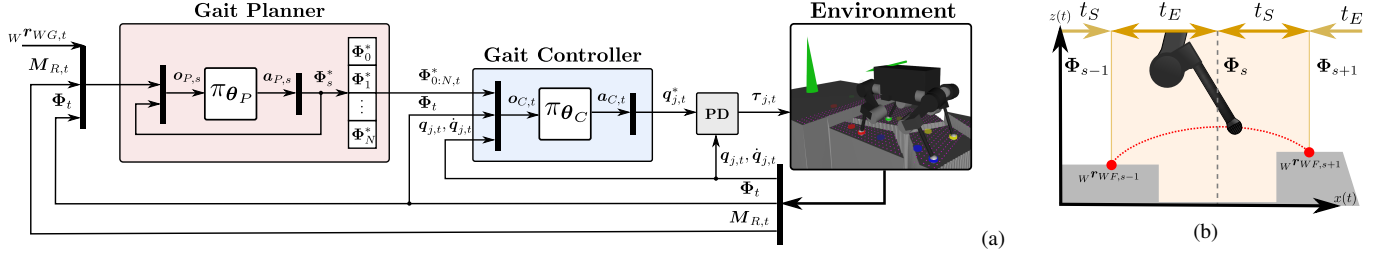


Fig. 2: (a) Overview of the proposed control structure used at deployment time. (b) Phases within a sequence are indexed using s , and every index corresponds to a point in time centered around a window defined by the durations t_E and t_S . The center of the window is defined by the motion of the base as captured by the phase Φ_s . t_S defines the time-to-switch from the current contact support to the next, specified in Φ_{s+1} , and t_E defines the time elapsed since the switch from the previous contact support, specified in Φ_{s-1} , to the current.

Support Phase Transition Feasibility: Transition feasibility amounts to evaluating if a feasible motion exists between a pair of support phases Φ_t, Φ_{t+1}^* , where the former is assumed while the latter is a candidate successor. As previously mentioned, we employ the general framework defined in [14] to design a convex LP using the Convex Resolution Of Centroidal dynamics trajectories (CROC) formulation. We use CROC to derive a set of linear equality and inequality constraints, forming a convex polytope, using the following terms:

- 1) A Centroidal Dynamics model of the system.
- 2) The contact force unilateral and friction constraints.
- 3) Assume angular momentum rate of zero.
- 4) Parameterization of CoM motion as a Bezier curve.
- 5) Restrict motion of the feet w.r.t. the base.
- 6) Restrict contact forces in magnitude and direction.

The specific variant of CROC we employ uses a trivial cost, a time-discretization of the CoM trajectory, and incorporates the parameterization of the contact forces into the decision variables of the optimization. We defer the reader to [14] for further details regarding CROC. The resulting formulation allows us to realize the transition feasibility mapping $F_{croc} : \Phi \times \Phi \rightarrow \{0, 1\}$. Therefore, we evaluate the LP for pairs of phases Φ_t, Φ_{t+1}^* , to determine if the corresponding phase transition is feasible, 1, or not, 0.

MDP Specification: Modeling locomotion as discrete sequences of support phases Φ , as defined in Sec. II-B, allows us to formulate phase transitions that exhibit the Markov property [15], and thus lends itself to modeling the overall problem of gait planning as an MDP. Specifically, we define the state of the MDP as the tuple $s_P := \langle \Phi, r_G \rangle$, where r_G is the current absolute position of the goal in the world, while observations and actions are defined as the tuples $o_P := \langle o_R, o_v, o_F, o_c, o_M \rangle$ and $a_P := \langle a_R, a_B, a_v, a_F, a_c, a_t \rangle$, respectively. Observations, consist of terms pertaining to the current state of the robot and the coincident terrain in the form of: the attitude w.r.t the goal $o_R \in \mathbb{R}$, the CoM velocity $o_v \in \mathbb{R}^2$, the feet positions $o_F \in \mathbb{R}^8$, the feet contact states $o_c \in \mathbb{R}^4$ and the local height-map $o_M \in \mathbb{R}^{32 \times 32}$. Conversely, actions, contain terms pertaining to changes to the current phase Φ , in the form of the CoM rotation $a_R \in \mathbb{R}_{clip}$, CoM translation $a_B \in \mathbb{R}_{clip}^2$, CoM velocity $a_v \in \mathbb{R}_{clip}^2$, feet positions $a_F \in \mathbb{R}_{clip}^8$, feet contact states $a_c \in \mathbb{R}_{clip}^3$ and the phase timings $a_t \in \mathbb{R}_{clip}^2$. All action terms are scaled, offset and clipped to lie in $\mathbb{R}_{clip} := [-1; 1]$. The exact definitions for actions and observations are provided in Tab. I.

TABLE I: Transformations used to form o_P from phases and phase transitions from a_P : the matrix $R_z(\alpha)$ defines a rotation about the world's z-axis by angle α , and $f_{dec} : \mathbb{R}_{clip}^3 \rightarrow \{0, 1\}^4$ decodes a 3-digit binary encoding into a vector of contact states.

GP Observation	GP Action
$o_v = B^v B_{x,y}$	$R_B^* = R_z\left(\frac{\pi}{8} a_R\right) R_B$
$o_F = B^r B F_{x,y} - B^r B N_{x,y}$	$B^r B_{x,y} = B^r B_{x,y} + 0.3 a_B$
$o_c = 2 c_F - \mathbf{1}_{4 \times 1}$	$B^v B_{x,y} = a_v, c_F^* = f_{dec}(a_c)$
$o_M = M_R$	$B^r B F_{x,y} = B^r B N_{x,y} + 0.3 a_F$
$o_R = -atan2\left(\frac{B^r B G_{x,y}}{B^r B G_{x,x}}\right)$	$[t_E^*, t_S]^T = \mathbf{1}_{2 \times 1} + 0.9 a_t$

Transition Dynamics: We define state transition dynamics for this MDP employing a formalism defining so-called *termination condition* functions $T(s_{P,t}, a_{P,t}, s_{P,t+1})$, which determine if an episode terminates. By formulating an episode termination as a transition into an absorbing terminal state, we can say that, an episode under this MDP, terminates whenever $s_{P,s+t} = s_{P,s}, \forall t > 0$. In this MDP, in particular, we employ the following termination conditions:

- 1) $T_{footholds}$: Checks for obstacles or gaps within the vicinity of each foothold using a fixed eight-point grid surrounding each foot.
- 2) T_{base} : Checks for collisions between base and terrain.
- 3) $T_{feasibility}$: Evaluates $F_{croc}(\Phi_t, \Phi_{t+1}^*)$.

Thus, each step of this MDP proceeds as follows: (1) Given a state $s_{P,t}$, the MDP computes the corresponding observation $o_{P,t}$, which is constructed according to the transformations in Tab. I, and is passed to the agent to select an appropriate action according to π_{θ_P} . (2) The selected action $a_{P,t}$, is used to compute the candidate phase Φ_{t+1}^* , again using the set of transformations defined in Tab. I. (3) The aforementioned terminations conditions are used to assert if the phase transition is feasible. This formulation therefore allows the agent to propose the phase transition directly, while the MDP only checks if it is feasible and otherwise terminate the episode.

In addition, we outline certain considerations regarding the computation of Φ_{t+1}^* . First, if a foot would be in contact for both phases Φ_t and Φ_{t+1}^* , the new foothold is reset to that of Φ_t since stance feet cannot change positions. Second, the z-coordinates of the feet and CoM positions are adjusted according to the height of the terrain at their respective locations, and for the CoM in particular, the height is set to a constant h_{com} above the lowest foothold. Lastly, to evaluate

the kinematic constraints of CROC, we need to infer the intermediate orientations. To this end, we use a first-order approximation of the angular velocity of the base by linearly interpolating between the starting and next base attitudes of each transition. Doing so also renders zero angular momentum between consecutive support phases, which is inline with assumptions made by CROC. One can thus envision an extension to CROC that includes angular momentum, which we intend to explore in future work.

Reward Function: We design a reward function which drives the agent to learn behaviors for reaching the goal position, facing the goal as much as possible, minimizing kinematic effort during phase transitions and inhibiting long stance phases. The final reward function is specified as the combination of multiplicative and additive terms

$$r_P(\mathbf{s}_{P,t}, \mathbf{s}_{P,t}, \mathbf{s}_{P,t+1}) := r_p \cdot r_h^2 \cdot r_k - r_c \quad (2)$$

where r_p rewards the agent for bringing the average foothold position closer towards the goal and penalizes moving it away, r_h penalizes the robot for not facing the goal, r_k penalizes for moving the feet away from the nominal footholds ${}_B\mathbf{r}_{NF}$ located beneath the shoulders and r_c penalizes for not lifting a foot over multiple steps, therefore promoting exploration and prevents the policy from getting stuck in the local optimum of remaining in a constant stance.

Furthermore, we would like to emphasize certain features of the multiplicative term in the above reward function. Specifically, this term results in a penalty that is small when r_p is small, i.e., beginning of training, and large when r_p is large, i.e., towards the end of training, thus resulting in a form of automatic scaling of the overall multiplicative term. We found that using these multiplicative rewards results in beneficial gradients throughout all iterations of training, as their values are ensured never to be too large as to hinder exploration, and never too small as to have negligible effect. Furthermore, as r_p is computed using the average position of footholds and not of the base, the agent is required to walk in order to maximize reward, as opposed to just merely leaning. The latter aspect is important, since leaning forward also inhibits the motion of the front legs, therefore making it much harder to walk. Tab. II details the aforementioned reward terms.

Policy Definition: We parameterize the GP's policy as a Gaussian distribution with a diagonal covariance matrix $\pi_{\theta_P}(\mathbf{a}|\mathbf{o}_{P,t}) := \mathcal{N}(\mathbf{a}|\boldsymbol{\mu}_{\theta_P}(\mathbf{o}_{P,t}), \boldsymbol{\sigma}_{\theta_P})$. The mean $\boldsymbol{\mu}_{\theta_P}(\mathbf{o}_{P,t})$ is output by a NN which inputs both exteroceptive and proprioceptive measurements into a series of NN layers, similar to those proposed in [13]. First, \mathbf{M}_R is input into three CNN layers, the output of which is subsequently input into one more fully-connected layer. The resulting latent output from the height-map is concatenated with the raw proprioceptive measurements, then fed into two more fully-connected layers with ReLU and tanh nonlinearities, and finally passed through a linear output layer. The standard-deviation parameters $\boldsymbol{\sigma}_{\theta_P}$ are realized by an additional layer that is independent of the observations and is used to drive exploration during training. Fig. 3(a) provides a graphical depiction of the NN model. Due to the inclusion of high-dimensional height-map data in the observations \mathbf{o}_P as well as the relatively large dimensionality

of the actions \mathbf{a}_P , we trained π_{θ_P} with a variant of Proximal Policy Optimization (PPO) using clipped loss and a Generalized Advantage Estimation (GAE) critic [16].

B. Gait Control

The GC is responsible for executing the support phase sequence provided by the GP while maintaining balance at all times. It operates by tracking a series of footholds and base positions extracted from the support phase sequence generated by the GP. In order to learn this behavior, we define an MDP with transition dynamics which incorporates the physics of the system and specify an appropriate parameterization for π_{θ_C} . Training a GC agent in such an MDP requires only that a target phase sequence be provided, and does not assume that a GP is available a priori. In fact, the target phase sequence can be provided arbitrarily as long as the target footholds are feasible. However, in this work, we elected to utilize a GP for this purpose solely as a matter of convenience so to avoid the use of additional elements.

Target Foothold Extraction: Assuming the GP is queried at some time t , we denote the resulting phase sequence as $\Phi_{0:N,t}^*$, where $\Phi_{0,t}^*$ is the initial phase as measured by the GP before generating the sequence of length N . This amounts to *rolling-out*² the planning policy by recursively evaluating π_{θ_P} using its own output.

MDP Definition: Given a phase sequence $\Phi_{1:N}^*$, the GC proceeds to extract the following target quantities: (a) target position for the base \mathbf{r}_B^* , (b) target feet contact states \mathbf{c}_F^* , and (c) target foothold positions \mathbf{r}_F^* for all legs. In the case of \mathbf{r}_F^* , targets are set by looking ahead into the phase plan so to ensure that both swing and stance legs have valid references at all times. Thus, the GC computes the foothold tracking errors ${}_B\mathbf{r}_{F,err}$, at 100Hz, while the target footholds are updated at approximately 2Hz. We specify MDP states \mathbf{s}_C , observations \mathbf{o}_C and actions \mathbf{a}_C defined as

$$\begin{aligned} \mathbf{s}_C &:= \langle \mathbf{R}_B, \mathbf{r}_B, \mathbf{v}_B, \boldsymbol{\omega}_B, \mathbf{q}_j, \dot{\mathbf{q}}_j, \mathbf{n}_F, \mathbf{c}_F \rangle \\ \mathbf{o}_C &:= \langle {}_B\mathbf{r}_{F,err}, \mathbf{c}_F^*, {}_B\mathbf{e}_z^W, z_{BF}, B\mathbf{v}_B, \\ &{}_B\boldsymbol{\omega}_B, \mathbf{c}_F, \mathbf{q}_j, \dot{\mathbf{q}}_j, \mathbf{q}_j^*, \eta \rangle, \quad \mathbf{a}_C := \langle \mathbf{q}_j^* \rangle \end{aligned} \quad (3)$$

where ${}_B\mathbf{e}_z^W$ is the gravity-aligned z-axis of frame W expressed in coordinates of frame B , z_{BF} is the distance between the lowest stance foot and the base along the z-axis of W , \mathbf{q}_j^* is the vector of previous target joint positions, and $\eta \in [0, 1]$ is a phase variable indicating the normalized time within a support phase. The transition dynamics of this MDP includes the generation of the phase plan using the GP, the physics of the system and the joint-space PD controller. As the PD controller is evaluated at 400Hz and the GC at 100Hz, we apply a zero-order hold of the joint positions output by the policy when computing joint torques commands. For this MDP, we also define the following termination conditions:

- 1) $T_{attitude}$: Angle between \mathbf{e}_z^B and \mathbf{e}_z^W exceeds 60° .
- 2) $T_{contact}$: Base collides with the terrain.

²The small range and dimensions selected for the elevation map, in conjunction with the limitation on maximum step length assumed by the planner, allows us to extract multiple successive samples of \mathbf{M}_R from within the effective field-of-view afforded by exteroceptive sensing.

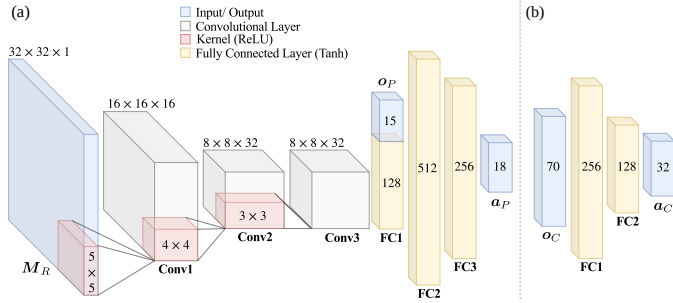


Fig. 3: The neural-network models used for the latent parameters of the (a) GP and (b) GC, respectively.

We have found that these two simple, yet effective, termination conditions are those principally responsible for the balancing and recovery behaviors learned during training, and thus their importance must not be understated. Moreover, we designed a reward function that emphasizes tracking of target foothold positions and contact states, but also contains terms that inhibit extraneous and aggressive motions during locomotion. The resulting reward function is defined as

$$r_C(\mathbf{s}_{C,t}, \mathbf{a}_{C,t}) := r_e + r_{tc} + r_{sw} + r_{sl} + r_t + r_v + r_a \quad (4)$$

where r_e and r_{tc} are the task-specific rewards penalizing deviations from the target foothold positions and contact states, r_{sl} penalizes foot-slip for grounded feet, r_{sw} penalizes large velocities for swing legs, r_t penalizes joint torques, r_v penalizes vertical linear and roll-pitch angular velocities of the base and r_a penalizes large angles between the unit vectors e_z^B and e_z^W of the base and world frame respectively.

Policy Definition: The GC’s policy, like that of the GP, is also parameterized as a Gaussian distribution with diagonal covariance matrix $\pi_{\theta_C}(\mathbf{a}|\mathbf{o}_{C,t}) := \mathcal{N}(\mathbf{a}|\boldsymbol{\mu}_{\theta_C}(\mathbf{o}_{C,t}), \boldsymbol{\sigma}_{\theta_C})$. While the mean $\boldsymbol{\mu}_{\theta_C}(\mathbf{o}_{C,t})$ is output by a simple NN with two fully-connected layers using *tanh* non-linearities, shown in Fig. 3 (b), the standard deviation coefficients $\boldsymbol{\sigma}_{\theta_C}$ are, just as in the case of the GP, output by an additional layer of parameters which is independent of $\mathbf{o}_{C,t}$. Due to the relatively small dimensionality of the MDP and π_{θ_C} , we train the latter using Trust-Region Policy Optimization (TRPO) also employing a GAE critic [17].

IV. RESULTS

A. Experimental Setup

In order to evaluate our approach, we crafted a suit of terrain scenarios for training and testing the GC and GP agents, as depicted in Fig. 1. The first and most basic scenario consists of an infinite flat plane we refer to as *Flat-World*, which we use to establish a baseline for performance and behavior. Secondly, the *Random-Stairs* terrain presents a $20 \times 20 \text{ m}^2$ square area consisting of $1 \times 1 \text{ m}^2$ flat regions of randomly selected elevation. The elevation changes were generated in a way that results in an effective inclination diagonally across the map. The third terrain scenario is that which we call *Temple-Ascent*, and is a composite terrain consisting of gaps, stepping stones, stairs as well as flat regions. We realized the MDP environment for the GP using an own implementation of CROC in C++, while for the MDP environment of the GC

we used the RaiSim [18] multi-body physics engine. All DRL algorithms were implemented using the *TensorFlow*³ C/C++ API⁴.

B. Gait Planner

Training Setup: Training of GP policies in the terrain suite consists of a set of episodes where the robot’s objective is to reach a goal position from a sufficiently distant starting location. Both starting and goal positions are selected randomly at the start of each episode. However, this procedure differs depending on the features of the terrain, as we must avoid invalid starting positions and unreachable goal positions, which, would negatively impact the resulting policies during training. Once valid starting and goal positions have been sampled, the robot’s initial attitude and footholds are also sampled uniformly from within respective bounds.

We thus trained two separate GP policies for *Random-Stairs* and *Temple-Ascent* respectively, using PPO with only 14 parallel workers running on the respective desktop computer over $200k$ iterations, which amounts to a total of two billion samples per run. Hyper-parameter values are listed in Tab. IV. We did not need to train a separate GP in *Flat-World*, and instead used that trained in *Temple-Ascent* for the respective performance evaluations.

Performance Metrics: In order to assess the performance of GP policies, we define the Episodic Success Rate (ESR), which measures the number of successfully reached goal positions over a finite number of episodes. Essentially, we execute a sufficiently large number of episodes where the robot tracks a reference goal position in the world and assert if the robot has reached within a 0.5 m vicinity of the goal and within a maximum permissible episode duration.

Training Results: GP training required approximately 82 hours in each terrain scenario. Throughout our experiments we found that the randomization scheme mentioned above and used for realizing the initial state distribution of the MDP was crucial for successfully learning to traverse all parts of the terrains. This demonstrates that if the agent does not observe all aspects of the terrain from the very beginning of training, it is often unable to generalize to unseen cases at test time.

Furthermore, we observed that, as the centroidal dynamics model employed by CROC is relatively conservative, it tends to limit the set of transitions that the policy can generate. This conservativeness is furthered by the fact that in this work, we limit the possible contact states that the GP’s policy can output to only those with three and four active contacts. Such a restriction was helpful for reducing the complexity of the problem, and we intend to extend to the general case of two and single contact configurations in future work.

We tested the GP policies in their respective terrain scenarios and evaluated their performance using the ESR metric. In all cases, we have observed that fully trained policies can generate valid support phase sequences which lead the robot

³<https://github.com/leggedrobotics/tensorflow-cpp>

⁴For the GP, we used a PC with 2x Intel Xeon E5-2680v4 (@2.4GHz) CPUs, 128GB of RAM, and an Nvidia GTX Titan (Pascal), and for the GC a PC with a single Intel Core i7-8700K (@3.7GHz) CPU, 64GB of RAM and an Nvidia GTX 2080 Ti GPU

TABLE II: The GP and GC reward terms. The subscript k indexes each foot, and $n_{contact,k}$ counts the current number of consecutive state updates for which foot k has remained grounded. b_k is a binary variable specifying whether a foot is within a cylinder of radius $d = 5$ cm of target foothold. An over-lined letter describes the conjugate of the binary variable, e.g. \bar{b}_k . The weighting factors are: $w_p = 25$, $w_k = 80$, $w_c = 0.01$, $w_{tc} = 0.1$, $w_{sw} = 0.01$, $w_t = 0.001$, $w_v = 0.5$, $w_e = 2$, $w_{sl} = 0.02$, $w_a = 0.2$.

Gait Planner Rewards	Gait Controller Rewards
$r_p = w_p \left\ \mathbf{r}_G - \sum_{k=1}^4 \mathbf{r}_{F,k} \mathbf{c}_{F,k} \right\ _2 - w_p \left\ \mathbf{r}_G - \sum_{k=1}^4 \mathbf{r}_{F,k}^* \mathbf{c}_{F,k}^* \right\ _2$	$r_{tc} = w_{tc} \sum_{k=1}^4 \left[\bar{b}_k (c_{F,k} - \bar{c}_{F,k}) + \bar{b}_k (c_{F,k} (c_{F,k}^* - \bar{c}_{F,k}^*) + \bar{c}_{F,k} (\bar{c}_{F,k}^* - c_{F,k}^*)) \right]$
$r_k = \max \left[0, 1 - w_k \left(\sum_{k=1}^4 \ \mathbf{r}_{NF,k,x}\ _2^3 + \ \mathbf{r}_{NF,k,y}\ _2^3 \right) \right]$	$r_{sw} = -w_{sw} \sum_{k=1}^4 \left[\bar{c}_{F,k} \ \mathbf{v}_{F,k}\ _2^2 - \bar{b}_k \bar{c}_{F,k}^* \min(d, \mathbf{r}_{F,k,z} - \mathbf{r}_{F,k,z}^*) \right]$
$r_h = 1 - \frac{1}{\pi} \left \text{atan2} \left(\frac{\mathbf{B} \mathbf{r}_{BG,y}}{\mathbf{B} \mathbf{r}_{BG,x}} \right) \right , r_c = w_c \sum_{k=1}^4 n_{c,k}$	$r_t = -w_t \ \boldsymbol{\tau}\ _2^2, r_a = -w_a \text{acos}((\mathbf{e}_z^B)^T \mathbf{e}_z^W), r_{sl} = -w_{sl} \sum_{k=1}^4 c_{F,k} \ \mathbf{v}_{F,k,xy}\ $
	$r_e = -w_e \sum_{k=1}^4 \sqrt{\ \mathbf{r}_{F,k}^* - \mathbf{r}_{F,k}\ _2}, r_v = -w_v \mathbf{B} \mathbf{v}_{B,z} ^2 - \ \mathbf{B} \boldsymbol{\omega}_{B,xy}\ _2^2$

to the goal with at an average ESR nearing 100.0%. The performance of GP policies trained and tested in the terrain suite are presented in Tab. III, where they have been deployed together with respective GC policies.

Another important observation regarding the output of the GP has to do with the types of gaits it manifests. In the case of *Flat-World* as well as in the flat regions of *Temple-Ascent*, we observe that the GP tends to output mostly cyclic support phases. This indicates that the agent learns to generate cyclic gaits even though no aspect of the MDP ever directed it to do so. In certain cases, however, such as the *Stepping-Stones* and *Gaps* bridges as well as when performing sharp point-turns, the GP outputs acyclic support phases.

Sample Complexity: One key contribution of this work is the significant reduction in sample complexity afforded by the use of transition feasibility instead of physical simulation to formulate the GP's MDP. Using the transition feasibility check, we can evaluate the MDP's transition dynamics at several thousands of steps-per-second, where each step corresponds to potentially several seconds of simulation time. Conversely, using a physics simulator typically requires several hundred or even thousands of steps to evaluate just one second of simulation time. Specifically, during training, we executed episodes with a maximum length of 50 steps with each corresponding to an average duration of approximately 2.6 s, which amounts to 130 s of simulation time. However, the physics simulator using a time-step of 2.5 ms would require $24k$ steps to simulate the duration above. As the throughput of the transition feasibility LP and the physics simulator, for our formulation, is $1k$ Hz and $60k$ Hz respectively, we can estimate an 18-fold effective reduction in sample-complexity.

C. Gait Controller

Training Setup: Training a GC agent involves collecting MDP transitions over a rich set of target footholds. In order to achieve such a distribution of training samples, we ensure that both the initial state distribution of the MDP as well as the target footholds generated by the GP are appropriately and sufficiently randomized. Initial states are generated by first uniformly sampling *initial* and *goal* positions of the base from within the bounds of the world. We then orientate the base by sampling uniformly from attitudes centered on the current orientation facing the goal, and bounded by the vector of Euler angles $[0.1, 0.1, \pi/4]$. Moreover, we randomize the initial feet positions by uniformly sampling xy coordinates from a $0.1 \times$

0.1 m^2 box defined in the base frame B and centered around the nominal values that would place the feet below the shoulders. Furthermore, to randomize the target footholds, we perform a randomized fixed rollout of the GP up to the maximum permissible length of an episode. Essentially, we rollout the GP however many times necessary such that the resulting phase sequence meets or exceeds the duration time of an episode, and randomize the target footholds at each step by adding a bias uniformly sampled from $[-0.1, 1.0]$ in the xy plane while ensuring that the z coordinates are fixed to the terrain.

With the aforementioned sampling scheme, we trained a GC agent using TRPO in *Flat-World* using only 24 parallel workers for total of $20k$ iterations. Moreover, as part of ongoing work to extend our method to full 3D foothold tracking, we present preliminary results for GC policies for stair-climbing by first pre-training in *Random-Stairs* then also on the stairs section of *Temple-Ascent*. Tab. IV presents the hyper-parameters most pertinent to the training of GC, for all of the cases mentioned above. We want to emphasize that in all cases, the same hyper-parameters were used, as we only adapted the initial state distribution accordingly for each terrain. TRPO was employed using mostly the default hyper-parameters specified in [17], [19].

Performance Metrics: We define two metrics for quantifying the performance of GC policies at test-time. First we define the *Foothold Tracking Error Rate* (FTER)

$$FTER := \frac{1}{T} \sum_{t=0}^T \frac{1}{\sum_{k=1}^4 c_{F,k}^*} \sum_{k=1}^4 c_{F,k}^* \|\mathbf{r}_{F,k}^* - \mathbf{r}_{F,k}\|_2 \quad (5)$$

that measures the mean foothold tracking error throughout an individual episode of length T , and is computed as a function of the desired contact states $c_{F,k}^*$, the desired foothold positions $\mathbf{r}_{F,k}^*$ and the measured feet positions $\mathbf{r}_{F,k}$ while in contact with the terrain for each foot and at every time-step. Secondly, we define the *Foothold Tracking Score* (FTS) as the ratio of successfully tracked footholds over the total generated by the GP within an episode. At the end of each support phase, we check if feet which were previously in swing phase have contacted the ground within 5 cm of the target foothold in the xy plane, and increment the FTS by one for each foot with a successful touchdown in the aforementioned region. These metrics are important with regard to the combined use of the GP and GC as they quantify how reliably a GC can execute the footholds generated by the GP. The planner has been trained to select footholds within a minimum distance of

TABLE III: Performance of the GC on the different terrain scenarios in *Temple-Ascent*, and under different kinds of variations to the system. The nominal system is that with which the GC was trained, and all variations are performed only at test time. m_B is the mass of the base, while l_{shank} is the length of the shank links. ESR values are listed as percentages, and all results are presented as empirical means plus-minus the corresponding standard deviations.

System	Metric	Flat	Gaps	Stepping-Stones	Stairs
Nominal	ESR	99.8% \pm 0.2%	96.4% \pm 2.3%	96.8% \pm 1.2%	90.6% \pm 6.8%
	FTS	0.985 \pm 0.000	0.967 \pm 0.000	0.970 \pm 0.000	0.751 \pm 0.000
	FTER	0.016 \pm 0.000	0.023 \pm 0.000	0.021 \pm 0.000	0.049 \pm 0.000
$m_B + 25\%$	ESR	99.4% \pm 0.8%	94.6% \pm 6.8%	98.4% \pm 0.8%	82.4% \pm 14.3%
	FTS	0.916 \pm 0.000	0.906 \pm 0.000	0.895 \pm 0.000	0.605 \pm 0.000
	FTER	0.028 \pm 0.000	7.332 \pm 266.3	0.032 \pm 0.000	0.060 \pm 0.000
$l_{shank} + 10\%$	ESR	99.0% \pm 1.5%	95.2% \pm 8.7%	97.6% \pm 0.3%	76.4% \pm 8.8%
	FTS	0.968 \pm 0.000	0.952 \pm 0.000	0.975 \pm 0.000	0.618 \pm 0.000
	FTER	0.020 \pm 0.000	0.025 \pm 0.000	0.020 \pm 0.000	0.069 \pm 0.000
$l_{shank} - 10\%$	ESR	100.0% \pm 0.0%	97.8% \pm 3.2%	97.4% \pm 0.8%	89.6% \pm 16.8%
	FTS	0.990 \pm 0.000	0.965 \pm 0.000	0.971 \pm 0.000	0.541 \pm 0.000
	FTER	0.017 \pm 0.000	0.022 \pm 0.000	0.021 \pm 0.000	0.058 \pm 0.000

5 cm from any changes in elevation exceeding 1 cm. As long as the controller can maintain foothold tracking within this region, then the combined system is ensured to operate safely.

Training Results: GC training endured for approximately 58 hours for *Flat-World* and approximately 116 hours for *Random-Stairs* and *Temple-Ascent*. The discrepancy in durations is due to the increased computational cost incurred in the physics engine when evaluating contacts between the terrain mesh and the multi-body system. Training in *Flat-World* results in a policy that succeeds in generalizing well to planar foothold tracking, while training in *Random-Stairs* and *Temple-Ascent* extends these capabilities to 3D. However, the stair-climbing agent trained in the latter case exhibits worse MER and FTS than those trained in the former. This difference is due to the difficulties in designing sampling schemes that always initialize the robot in valid initial states, but also to the similarity of the foothold targets generated by the GP as a result of the repetitive terrain features in the suite.

We evaluated the performance of the GC policies within *Temple-Ascent* across five runs, each consisting of 100 episodes with a maximum length of 90 s. We also perturb the model of the robot (i.e., with which the GC was trained) to assess the robustness of the policies. Specifically, we increased the mass of the base by 25% and varied the lengths of the shank links by $\pm 10\%$. In each case, ESR, FTS, and FTER values were recorded in order to compute empirical means and standard deviations. The resulting performance measurements are presented in Tab. III⁵ and Fig. 4 shows time-series plots of the policy overcoming a large 40 cm gap. The final policies deployed are demonstrated in the supplementary video⁶.

D. Comparison to Existing Approaches

As mentioned in Sec. I, most methods addressing planning and control of multi-contact motions employ optimization, sampling-based search, or a combination thereof. Although those relying solely on optimization [2]–[4] can be kinodynamic, they are not feasible online, yet those that employ

⁵Although mean performance for stairs is $\geq 90\%$ in the nominal case, the variance is noticeably higher for the perturbed models, indicating that the policy is more sensitive w.r.t model variations than that for other terrains.

⁶<https://youtu.be/s1rrM1ocZl4>

TABLE IV: Policy optimization algorithm hyper-parameters for the GC using TRPO and the GP using PPO (see [16], [19] for details).

Parameter	Symbol	TRPO	PPO
Batch Size	N_B	24k	200k
Mini-Batches	N_{MB}	-	5
Max. Episode Length	T_{max}	3000	50
Discount Factor	γ	0.995	0.99
Trace Decay	λ	0.99	0.97
Terminal Reward	r_T	-5.0	-1.0
KL Constraint	δ	0.01	-
Clip	ϵ	-	0.2
Entropy Weight	β	0.001	0.004
Initial Variance	σ_0^2	0.4	1.0
Adam Epochs	n_{epoch}	-	3
Adam Learning-Rate	α_{Adam}	-	0.0002
Gradient Clipping	g_{max}	-	1.0
CG Damping	β_{CG}	0.1	-
CG Steps	n_{CG}	40	-

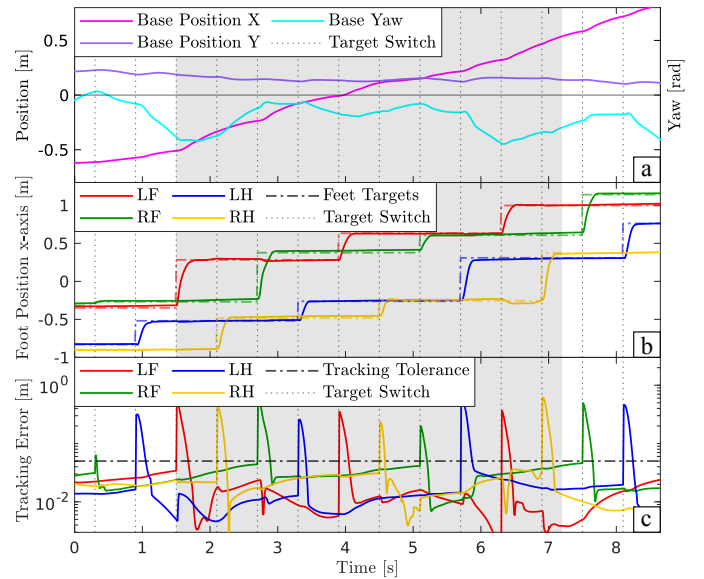


Fig. 4: Plots of the GC policy overcoming the 40 cm gap: (a) planar pose of the base, (b) desired vs measured positions of the feet, and (c) log-scale norm of the foothold tracking errors in the X-Y plane. The vertical dotted lines denote the times at which the foothold targets change according to the phase plan, gray regions denote the time spent crossing the gap, and the horizontal line in (c) denotes the 5 cm error tolerance defined by the foothold tracking cylinder.

sampling-based search [7], [8], or both [5], [6], [9] can be used online but remain kinostatic. The issue is that kinostatic methods are unable to fully exploit the dynamics of the system, and typically decouple foothold selection from the optimization of the base and feet motions, limiting their possible set of solutions. In addition, nearly all of the aforementioned approaches use some form of modelling [2], [6] or qualification [4], [5], [7] of the terrain. On the contrary, our approach relies on a minimal geometric representation of the terrain in the form of a height-map, performs gait-free kinodynamic planning by jointly optimizing base and swing feet motions together with the selection of footholds and is executable online.

Lastly, to illustrate how our method performs compared to others, we present an experiment in which the robot is to walk over a set of large gaps with different sizes.

Specifically, we compare with *Free-Gait*, a state-of-the-art model-based framework for perceptive locomotion [4] which jointly optimizes body poses and footholds using an Sequential Quadratic Programming solver. The experiment consists of two trials, corresponding to gaps of different sizes: 30, 40 cm. Fig. 5 provides snap-shots of the respective trials. Only our policies were able to traverse the 40 cm gap. *Free-Gait* relies on a heuristic scoring of the surrounding terrain and selects footholds using a deterministic greedy search within the vicinity of a nominal foothold. As it ensures that the CoM remains within some margin of the support polygon, it does not sacrifice static stability, even momentarily, to extend the reach of the front legs. Conversely, our planner can identify valid footholds across the gap in a single shot, and our controller then orientates the base to extend leg reach while retaining balance at all times.

V. CONCLUSION & DISCUSSION

This work proposes a new approach for training a two-layer hierarchy of NN policies which realizes terrain-aware locomotion for quadruped robots. We decompose locomotion into two parts trained independently using model-free DRL, and which interface via a carefully designed parameterization of quadrupedal gaits. As physical simulation incurs a high computational cost for use with DRL, our method reduces sample-complexity for training the gait planner by using a *transition feasibility* criteria realized as convex LP to formulate the transition dynamics of an MDP. Within a certain context, our formulation can resemble bilevel optimization schemes used in nonlinear programming. The upper-level optimization performed by model-free DRL optimizes the selection of footholds and instantaneous CoM motion for the coincident terrain. The lower-level optimization performed by the model-based LP optimizes phase transitions, thus resulting in optimal CoM trajectories between the support phases proposed by the higher-level. Although in this work we only used a trivial cost for the LP, we could instead formulate a cost that penalizes contact forces and CoM accelerations, and include it as an additional reward term in the MDP. Such a coupling of the LP and MDP optimization objectives would allow us to reason about both feasibility and optimality of the resulting gait plan. Furthermore, as the planner learns to generate relatively conservative foothold sequences due to the restrained centroidal dynamics model, the burden of realizing fully dynamic and optimal motions for the base and swing-feet is placed on the gait controller. Lastly, we have demonstrated the efficacy of this approach by successfully training and testing in a suite of challenging terrain scenarios. The resulting policies not only prove to be effective in the suite of rigid non-flat terrains but also manage to generalize well to previously unseen cases.

REFERENCES

- [1] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch, *et al.*, “ANYmal-a highly mobile and dynamic quadrupedal robot,” in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2016, pp. 38–44.
- [2] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake, “Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot,” *Autonomous Robots*, pp. 429–455, 2016.

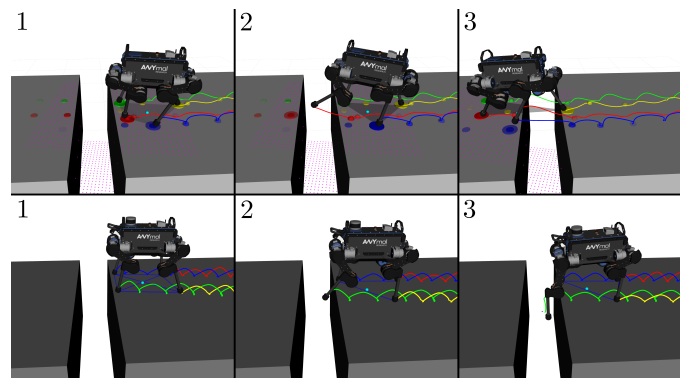


Fig. 5: Snap-shots of the comparison between our policies (top row) and *Free-Gait* (bottom row) overcoming the 40 cm gap. The figures are numbered left-to-right to indicate the order of the frames.

- [3] A. W. Winkler, C. D. Bellicoso, M. Hutter, and J. Buchli, “Gait and trajectory optimization for legged systems through phase-based end-effector parameterization,” *IEEE Robotics and Automation Letters*, pp. 1560–1567, 2018.
- [4] P. Fankhauser, M. Bjelonic, C. D. Bellicoso, T. Miki, and M. Hutter, “Robust rough-terrain locomotion with a quadrupedal robot,” in *IEEE Int. Conf. on Robotics and Automation*. IEEE, 2018, pp. 1–8.
- [5] D. Belter, P. Labecki, and P. Skrzypczyński, “Adaptive motion planning for autonomous rough terrain traversal with a walking robot,” *Journal of Field Robotics*, pp. 337–370, 2016.
- [6] M. Kalakrishnan, J. Buchli, P. Pastor, M. Mistry, and S. Schaal, “Learning, planning, and control for quadruped locomotion over challenging terrain,” *Int. Journal of Robotics Research*, pp. 236–258, 2011.
- [7] S. Tonneau, N. Mansard, C. Park, D. Manocha, F. Multon, and J. Pettré, “A reachability-based planner for sequences of acyclic contacts in cluttered environments,” in *Robotics Research*. Springer, 2018, pp. 287–303.
- [8] S. Tonneau, A. Del Prete, J. Pettré, C. Park, D. Manocha, and N. Mansard, “An efficient acyclic contact planner for multiped robots,” *IEEE Transactions on Robotics*, vol. 34, no. 3, pp. 586–601, 2018.
- [9] R. J. Griffin, G. Wiedebach, S. McCrory, S. Bertrand, I. Lee, and J. Pratt, “Footstep planning for autonomous walking over rough terrain,” *arXiv preprint arXiv:1907.08673*, 2019.
- [10] O. A. V. Magaña, V. Barasuol, M. Camurri, L. Franceschi, M. Focchi, M. Pontil, D. G. Caldwell, and C. Semini, “Fast and continuous foothold adaptation for dynamic locomotion through CNNs,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2140–2147, 2019.
- [11] T. Klamt and S. Behnke, “Towards learning abstract representations for locomotion planning in high-dimensional state spaces,” *arXiv preprint arXiv:1903.02308*, 2019.
- [12] N. Heess, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami, M. Riedmiller, and D. Silver, “Emergence of locomotion behaviours in rich environments,” *arXiv preprint arXiv:1707.02286*, 2017.
- [13] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, “DeepLoco: Dynamic Locomotion Skills Using Hierarchical Deep Reinforcement Learning,” *ACM Trans. Graph.*, pp. 41:1–41:13, 2017.
- [14] P. Fernbach, S. Tonneau, O. Stasse, J. Carpentier, and M. Taix, “C-CROC: Continuous and Convex Resolution of Centroidal Dynamic Trajectories for Legged Robots in Multicontact Scenarios,” *IEEE Transactions on Robotics*, 2020.
- [15] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [17] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” *arXiv preprint arXiv:1506.02438*, 2015.
- [18] J. Hwangbo, J. Lee, and M. Hutter, “Per-contact iteration method for solving contact dynamics,” *IEEE Robotics and Automation Letters*, pp. 895–902, 2018.
- [19] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust Region Policy Optimization,” in *Int. Conf. on Machine Learning*, 2015, pp. 1889–1897.