

# DeepWalk: Omnidirectional Bipedal Gait by Deep Reinforcement Learning

Diego Rodriguez and Sven Behnke

**Abstract**—Bipedal walking is one of the most difficult but exciting challenges in robotics. The difficulties arise from the complexity of high-dimensional dynamics, sensing and actuation limitations combined with real-time and computational constraints. Deep Reinforcement Learning (DRL) holds the promise to address these issues by fully exploiting the robot dynamics with minimal craftsmanship. In this paper, we propose a novel DRL approach that enables an agent to learn omnidirectional locomotion for humanoid (bipedal) robots. Notably, the locomotion behaviors are accomplished by a single control policy (a single neural network). We achieve this by introducing a new curriculum learning method that gradually increases the task difficulty by scheduling target velocities. In addition, our method does not require reference motions which facilitates its application to robots with different kinematics, and reduces the overall complexity. Finally, different strategies for sim-to-real transfer are presented which allow us to transfer the learned policy to a real humanoid robot.

## I. INTRODUCTION

Humanoid robots are one of the most versatile and flexible platforms for acting in made-for-human environments. This versatility comes, however, at the cost of complexity. Bipedal locomotion poses still several challenges in terms of planning and control mainly due to the high dimensionality, complex dynamics, sensing and actuation limitations, and real-time constraints. Inspired by the human example, Deep Reinforcement Learning (DRL) approaches offer a promising model-free alternative to address these issues by making use of prior experiences.

Several state-of-the-art DRL-based locomotion controllers employ tracking-based policies of reference motions to learn separate controllers to walk at fixed velocities [1]–[4]. In contrast, we propose a single policy in order to avoid training and combining separate control policies. The parameter sharing facilitates information transfer when learning different velocities. We further reduce complexity by removing the need for motion capture data and engineered modules such as kinematic mappings and motion interpolations. This is made possible by the introduction of a nominal pose that can be defined as the standing pose of the robot.

The core contribution of this paper is a novel approach to learn a *single* control policy for *omnidirectional* walking using DRL on a realistic humanoid robot model without using reference motions. To achieve this, we propose a velocity scheduler that gradually increases the task difficulty of the agent, and we introduce a nominal pose to guide the learning process. We additionally introduce a new way of

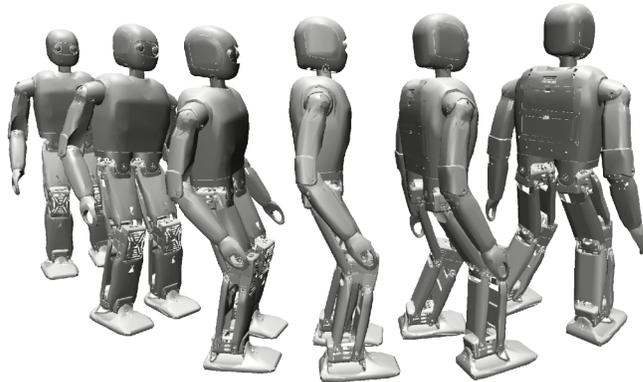


Fig. 1. Omnidirectional walk executed by our learned control policy. The robot is able to turn left during forward walking.

controlling the use of motor power (torque) by bounding the action space through beta policies [5].

Our experiments demonstrate that the learned policy can successfully produce omnidirectional motions allowing the robot to walk forwards, backwards, laterally, diagonally, to turn around the vertical axis and to combine these directions. Moreover, our learned controller is successfully transferred to real hardware.

## II. RELATED WORK

Learning approaches applied to walking controllers have been predominantly investigated for gait optimization [6]–[8]. These methods, however, depend on engineered components such as trajectory planners, central pattern generators and robot dynamics models. Recently, model-free locomotion controllers have been developed by the character animation community based on DRL algorithms [1], [9]–[13]. Heess *et al.* [9] generated robust locomotion maneuvers for different characters by applying curriculum strategies on the environment. In [10], 3D walking motions were achieved by employing curriculum strategies on the commanded velocities. Still, the policies were trained on animation characters and their applicability in real robots is questionable.

Peng *et al.* [1] provide human motion capture data with desired foot-placement goals to guide the reinforcement learning process. Peng *et al.* [11] later proposed DeepMimic, an approach that is able to mimic highly dynamic motions such as backflips, cartwheels, and rolls. Inspired by DeepMimic, Bergamin *et al.* [12] and Park *et al.* [13] learned kinematic trajectories which are tracked by dynamic-consistent RL-based tracking controllers. However, in all these approaches, the controllers were demonstrated on simplified models with

high torque capabilities and ground truth data from the simulator which do not resemble real-world conditions.

Deep reinforcement learning has also been applied in the robotics community [2], [3], [14]–[17]. Hwangbo *et al.* [14] presented a control policy which is later transferred to a real quadruped robot. The transfer was possible thanks to a network trained to model the dynamics of the actuators. In contrast to [14], our approach employs a much simpler reward function and it is demonstrated in a humanoid robot which imposes a harder balance problem.

One of the first attempts to apply DRL on humanoid robots was proposed by Yang *et al.* [15], who integrated the capture-step equations into the reward function for learning push-recovery capabilities. However, walk capabilities were not developed. Xie *et al.* [2] learned a forward walk with a biped robot using DRL based on separate reference motions at different velocities. Interpolation between policies needs to be explicitly handled to allow the robot to change between commanded velocities. Later, Xie *et al.* [3] extended this approach and transferred the policies to a real platform. Similarly, [17] transferred a walking controller into a real bipedal robot. In both approaches, the achieved locomotion capabilities rely on multiple control policies that need to be trained separately. Our approach, on the other hand, is able to generate an omnidirectional gait using a single policy without any reference motion.

### III. BACKGROUND

#### A. Deep Reinforcement Learning

Reinforcement learning algorithms aim to find a policy  $\pi$  based on experiences that will guide an agent to solve a specific task. This task is modeled as a Markov Decision Process (MDP) defined by the tuple  $\{S, \mathcal{A}, P, \gamma, r\}$ , where  $S \in \mathbb{R}^n$  represents the state space,  $\mathcal{A} \in \mathbb{R}^m$  is the set of actions the agent can take,  $P : S \times \mathcal{A} \mapsto S$  models the dynamics of the system,  $\gamma \in [0, 1]$  is a discount factor, and  $r : S \times \mathcal{A} \mapsto \mathbb{R}$  is a reward function that rewards or punishes an action  $a_t$  taken in state  $s_t$  after interacting with the environment at time step  $t$ .

In this paper, we focus on model-free reinforcement learning, and we directly construct a parametrized policy  $\pi_\theta$  by maximizing a cost function  $J(\theta)$  with respect to the parameters  $\theta$ , without explicitly modeling the dynamics  $P$ . As common in continuous control problems, we define  $\pi_\theta$  as a stochastic policy  $\pi_\theta(a_t|s)$  defined as a probability distribution of taking an action  $a_t$  given a state  $s_t$ . Policy gradient algorithms try to solve this optimization problem by sampling trajectories around the current policy  $\pi_\theta$  and by updating the parameters  $\theta$  according to the gradient  $\nabla_\theta J(\theta)$  in an ascent fashion. This gradient is expressed as:

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}[\Psi] = \mathbb{E}[\Psi_t \nabla_\theta \log \pi_\theta(a_t|s_t)]. \quad (1)$$

Eq. (1) tells us how the parameters  $\theta$  should be updated judged by the score function  $\Psi_t$  which can take several forms including: the state-action value function,  $Q^\pi(s_t, a_t)$ ,

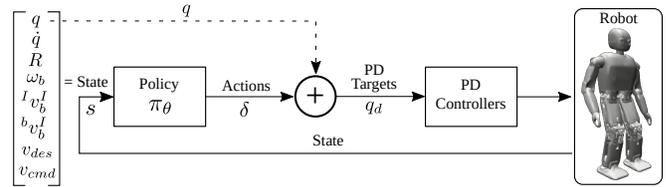


Fig. 2. Control system overview. According to the state  $s$  of the robot, the control policy  $\pi_\theta$  calculates increments  $\delta$  of the current joint state positions  $q$  that define targets  $q_d$  for PD controllers of the robot.

the advantage function,  $A^\pi(s_t, a_t)$ , and the Generalized Advantage Estimator (GAE):

$$A^{GAE(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}^V, \quad (2)$$

that reduces the variance of the gradient estimates at the cost of introducing bias [18]. This is done by combining a series of Temporal Difference (TD) residuals  $\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$  through a parameter  $\lambda \in [0, 1]$  that trades off variance and bias. The value function is defined as  $V^\pi(s_t) = \mathbb{E}_{s_{t+1}, a_t} [\sum_{l=0}^{\infty} r_{t+l}]$ .

Estimators of the gradient policy can additionally be obtained by automatic differentiation of an objective function constructed such that its gradient is the policy gradient estimator, e.g.:

$$L(\theta) = \mathbb{E}_t [A_t \log \pi_\theta(a_t|s_t)]. \quad (3)$$

Alternatively, a surrogate objective function can be used, as introduced in the Proximal Policy Optimization (PPO) algorithm [19]. This surrogate has the form:

$$L^{PPO}(\theta) = \mathbb{E}_t [\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t)], \quad (4)$$

where,  $r_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$  denotes the probability ratio, and  $\epsilon$  is a hyperparameter that defines a range in which the new policy is allowed to differ from the previous one.

### IV. METHOD

We propose a novel approach to learn an omnidirectional walking controller for humanoid robots. This controller is parametrized by a neural network called the policy (actor) network (Sec. IV-D). An overview of the walking controller is presented in Fig. 2. Our approach consists of two phases: training and inference. In the training phase, two networks (actor and critic) are learned through experiences that the agent collects acting in the environment using the current parameters of the networks. Once enough experiences have been collected, the networks' weights are updated and the experience rollout buffers are cleared out. In the inference phase, the network weights are kept fixed. For brevity, the dependency on  $t$  will be dropped. Given the observed state  $s$  of the robot, the network outputs offsets  $\delta$  that are added to the current joint positions  $q$  that ultimately define targets  $q_d$  for PD controllers of the robot joints.

### A. State Space

We define the state  $s$  of the robot as: joint positions  $q$ , joint velocities  $\dot{q}$ , orientation of the base of the robot  $R$ , angular velocity of the base  $\omega_b$ , velocity of the base w.r.t. a inertial reference frame  $I$  expressed in  $I v_b^I$ <sup>1</sup> and in the base link frame  ${}^b v_b^I$ , long-term desired velocity  $v_{des}$ , and short-term commanded velocity  $v_{cmd}$ .

The joint positions  $q \in \mathbb{R}^{n_q}$  and velocities  $\dot{q} \in \mathbb{R}^{n_q}$  are read directly from the joint encoders, for a robot with  $n_q$  fully actuated joints.  $R$  is a vector containing the orientation of the base link (roll  $R_\alpha$ , pitch,  $R_\beta$  and yaw  $R_\gamma$ ). Robots without yaw estimation will define a two-dimensional  $R$  vector.

The angular velocity  $\omega_b = [\omega_x, \omega_y, \omega_z] \in \mathbb{R}^3$  is taken from the gyro measurements. The linear velocity of the base link  $v_b$  is estimated using the robot kinematics and orientation assuming a flat ground. The reference frame  $I$  initially equals the base link frame and is updated periodically to a frame placed in the base link aligned with the  $z$  world axis. The linear velocity is expressed in the reference frame  $I v_b^I \in \mathbb{R}^2$  and in the local frame  ${}^b v_b^I \in \mathbb{R}^2$ , and both are included into the state  $s$ . Without  $I v_b^I$  in  $s$ , the agent would not be penalized for deviating from the original direction of the desired velocity; and without  ${}^b v_b^I$ , the agent is susceptible to develop locomotion patterns that follow the walking direction, but infringe the desired relative velocity.

Finally, the state also incorporates the desired velocity  $v_{des} = [v_x^{des}, v_y^{des}, \omega_z^{des}] \in \mathbb{R}^3$  and an immediately commanded velocity  $v_{cmd} = [v_x^{cmd}, v_y^{cmd}, \omega_z^{cmd}] \in \mathbb{R}^3$ . The former can be seen as the task's goal. It represents the user input for controlling the robot. The latter is an interpolated velocity between the previous desired velocity  $v_{des}^{old}$  and the new desired velocity  $v_{des}$  to have smoother transitions when changing velocities.

### B. Action Space

The actions are represented as a delta  $\delta$  that needs to be added to the current joint position  $q$ , such that PD targets are formulated as  $q_d = q + \delta$ . This formulation is based on the observation that abrupt changes in the PD target saturate the PD controllers producing jerky, unnatural high-torque motions. This issue is addressed by putting limits on the values of  $\delta$ , i.e., we respect actuator speed limits. In stochastic continuous control, the commonly used Gaussian policies are not able to handle bounded action spaces like the one proposed here. Chou *et al.* [5] introduced Beta policies, which can deal with bounded action spaces and have shown faster convergence and higher scores than the Gaussian ones. In our method, we follow Beta policies and bound the action space such that  $\delta \in [\delta_{min}, \delta_{max}]$ . The actions are sampled from the beta distribution at 40 Hz.

### C. Reward Function

The reward function is defined as:

$$r = w_v r_{vel} + w_r r_{reg} + w_a r_{alive} + w_f r_{foot}. \quad (5)$$

<sup>1</sup> Notation: for a vector  ${}^A v_B$ , the left superscript denotes that the coordinates of the vector are expressed in frame  $A$ , for the position of a point  $C$  relative to other point  $B$ .

All the reward terms are bounded such that  $r_* \in [0, 1] \cdot \Delta t$ , where  $\Delta t$  is the time step of the policy controller. To bound the reward terms, we use the smooth logistic kernel function  $K: \mathbb{R} \rightarrow [0, 1]$  expressed as  $K(x|l) = 2/(e^{lx} + e^{-lx})$ , where  $l$  defines the sensitivity of the kernel. The reward consists of the following terms:

1) *Velocity Tracking*: This term states how good the velocity is tracked by the control policy.  $r_{vel}$  is defined as:

$$r_{vel} = C_v K(e_v) \Delta t, \quad e_v = \left\| \begin{bmatrix} v_x^{cmd} - {}^b v_{b,x}^I \\ v_y^{cmd} - {}^b v_{b,y}^I \\ v_x^{cmd} - I v_{b,x}^I \\ v_y^{cmd} - I v_{b,y}^I \\ \omega_z^{cmd} - \omega_z \end{bmatrix} \right\|. \quad (6)$$

The tracking error  $e_v \in \mathbb{R}^5$  includes the difference between the commanded velocity  $v^{cmd}$  and the velocity of the base of the robot, expressed in the reference  $I v_b^I$  and the local frame  ${}^b v_b^I$ , and the angular velocity error  $\omega_z^{cmd} - \omega_z$ . Note that the L2-norm encourages improvement of all components together. The value of  $C_v \in (0, 1]$  changes dynamically the priority of this reward term. At the beginning of training, the priority is low to let the agent learn to stand. The priority is increased rapidly once the robot has a notion of standing. The value of  $C_v$  is defined per epoch:  $C_{t+1} = C_t^{k_d}$ , where  $k_d$  specifies the speed change of  $C_v$ . Without  $C_v$ , the agent would learn a greedy policy in which it tilts to the front, provoking a fall.

2) *Pose Regularization*: The regularization of the learning process in our approach is done through a single joint position configuration. This nominal pose can be defined, for example, as the standing pose of the robot. The regularization error is defined as the difference between the nominal pose  $q^{req}$  and the corresponding joint positions to regularize  $q^r$ :

$$r_{reg} = K(\|q^{req} - q^r\|) \Delta t. \quad (7)$$

Note that not all joints in  $q$  need to be regularized.

3) *Alive*: The agent is referred to be alive if it is not in the process of falling. In each step, a fixed reward is given if the height of the base with respect to the floor,  $p_z$ , remains above a defined value,  $p_z^{min}$ , or if the roll  $R_\alpha$  and pitch  $R_\beta$  angles of the base stay below thresholds,  $R_\alpha^{max}$ , and  $R_\beta^{max}$ , respectively. In case these thresholds are violated, a fall is expected and the rollout is terminated. In contrast to a large negative reward given at the terminal state,  $r_{alive}$  is proportional with the length of the rollout such that longer sequences are more rewarded, i.e., the agent is encouraged to learn keeping balance.  $r_{alive}$  is formulated as:

$$r_{alive} = \begin{cases} \Delta t & \text{if } p_z > p_z^{min}, R_\alpha < R_\alpha^{max}, R_\beta < R_\beta^{max} \\ 0 & \text{else.} \end{cases} \quad (8)$$

4) *Foot Clearance*: Without a foot clearance term, policies might be learned which lift the feet as little as possible. Although such policies show stable walking patterns in simulation, transferred policies to the real robot exhibit, in general, motions with dragging feet partly caused by model

differences and joint backlash. The foot clearance term acts on the swing leg only. Apart from considering the clearance  ${}^{lf,rf}p_z$ , this term includes the roll angle  ${}^{rf,lf}\phi_x$  of the right ( $rf$ ) and left ( $lf$ ) foot to discourage the agent to walk on its lateral feet edges which emerges as an artifact from maximizing  ${}^{lf,rf}p_z$ . For a right swing leg, for instance, the foot clearance reward term is formulated as:

$$r_{foot} = C_f K(e_f) \Delta t, \\ e_f = \left\| [w_\phi ({}^{rf}p_z^{des} - {}^{rf}p_z), {}^{rf}\phi_x, {}^{lf}\phi_x, w_\phi {}^{lf}p_z]^T \right\|. \quad (9)$$

The minimization of the clearance of the foot in stance disfavors the development of flight phases, which produce unstable behaviors when the policy is executed on the real robot. For a left swing leg, the superscripts of Eq. (9) are interchanged correspondingly. The swing leg is defined as the leg whose foot has the smallest distance to the trunk link in the  $z$  axis. Hysteresis is added when changing the swing leg to discourage flight phases.

#### D. Actor and Critic Networks

In our approach, two networks are trained: a critic and an actor network. The critic network is employed to estimate a state value function  $V_\phi$  used for calculating the generalized advantage estimator  $A^{GAE}$  (Eq. (2)). The parameters of the critic network are updated by minimizing the loss function:

$$L^V = (V_\phi(s_t) - \hat{V})^2, \quad (10)$$

where  $\hat{V}$  refers to the sampled state value from the trajectories. The parameters of the actor network are updated by maximizing the PPO loss function (Eq. 4).

The critic and actor network architectures are very similar. They contain two fully-connected hidden layers of 512 units with tanh activation functions and a fully-connected input layer for the state vector  $s_t$ . The last layer of the critic network contains a unit that represents the estimated state value  $V_\phi$ . The actor network has in its last layer two units for each dimension of the action space. These units parametrize the beta distributions of the actions. In training, the actions are sampled from these distributions whereas in inference the actions are described by the distribution mode. We train our control policy using PPO combined with GAE.

#### E. Curriculum Learning for Target Velocities

We formulate the problem of learning a control policy for omnidirectional walking as a curriculum learning problem [20]. This curriculum strategy is implemented as a velocity scheduler that increases the task difficulty gradually. The velocity scheduler defines the bounds from where a target velocity  $v^{des} \in \mathbb{R}^3$  is sampled each episode from a three-dimensional uniform distribution. In the first episode, the robot is commanded to walk only in the sagittal direction at a fixed velocity  $v_{core}$ . The bounds of the regions, from where the target velocities are sampled, are gradually increased as training progresses. We define an episode number  $\zeta$ , in which the target velocities stop increasing. The bound values for each episode are linearly interpolated according to  $\zeta$  and the

maximum target velocities. The training process continues until the maximum number of episodes is reached. In this manner, the agent can refine the learned policy.

The core velocity  $v_{core} \gg [0, 0, 0]$  is introduced for several reasons: first, to encourage more walking than standing behaviors, second, to avoid forcing the agent to learn at the beginning to walk forward and backward simultaneously, which is a harder task compared to walking forward, and finally, to prevent the agent learning to slide instead of walking.

#### F. Sim-to-real Transfer

In this section, we introduce several strategies that facilitate the sim-to-real transfer.

1) *System Identification*: This contributes to finding a good initial set of simulation parameters. We found that the reflected inertia is a decisive parameter for learning a stable gait. Specifically, we notice that low values of this parameter lead to jerky motions that are very challenging to control.

Tuning the PD controllers of the real robot and the simulation to get similar responses is critical. Because the responses of the PD controllers influence the joint state configuration and therefore the network input, a single untuned PD controller suffices to produce a previously non-observed input that might lead to continually increasing instabilities.

2) *Noise Injection*: Additional noise is purposely injected to the output network. At the beginning of each episode, noise is independently sampled for each joint. The noise is sampled from a uniform distribution  $\eta_{pd} \sim \mathcal{U}(1 - \epsilon_{pd}, 1 + \epsilon_{pd})$ .  $\eta_{pd}$  acts as a scale factor of the inferred target deltas  $\delta$ . The noised PD targets are then defined as  $q_d = q + \eta_{pd} \cdot \delta$ . To model real sensory data, noise is also added to the sensors (gyro  $\eta_g$ , accelerometer  $\eta_a$ , position encoders  $\eta_q$  and velocity sensors  $\eta_v$ ). The added sensory noises are independently randomly sampled from zero-mean Gaussian distributions.

3) *Dynamics Randomization*: Parameters such as masses, inertia and joint positions are obtained directly from the CAD model. For the sim-to-real transfer, the friction model plays a key role, and it needs to consider tangential and torsional forces. The friction between two surfaces is modeled by elliptic cones. The friction coefficients (tangential  $\mu_t$  and torsional  $\mu_z$ ) are randomly sampled from uniform distributions  $\mu_{t,z} \sim \mathcal{U}(\mu_{t,z}^{min}, \mu_{t,z}^{max})$  at the beginning of each episode.

4) *Modeling of Actuation Latency*: We define latency as the time an actuator takes to read its actual position after a commanded target has been sent. In simulation, this happens immediately. However, with the real robot, this delay implies that the learned policy is taking actions on states that do not fully represent the actual state of the robot, which leads to unstable gaits and jerky motions. Following the approach proposed by Tan *et al.* [4], we model actuation latency by keeping a history of observations and by feeding previous observations to the network according to a delay time  $t_{lat}$ . Observations are recorded at the frequency of the control policy. The network input is then defined as a linear interpolation between adjacent observations according to the latency  $t_i \leq t_{lat} \leq t_{i+1}$ . To make the policy robust

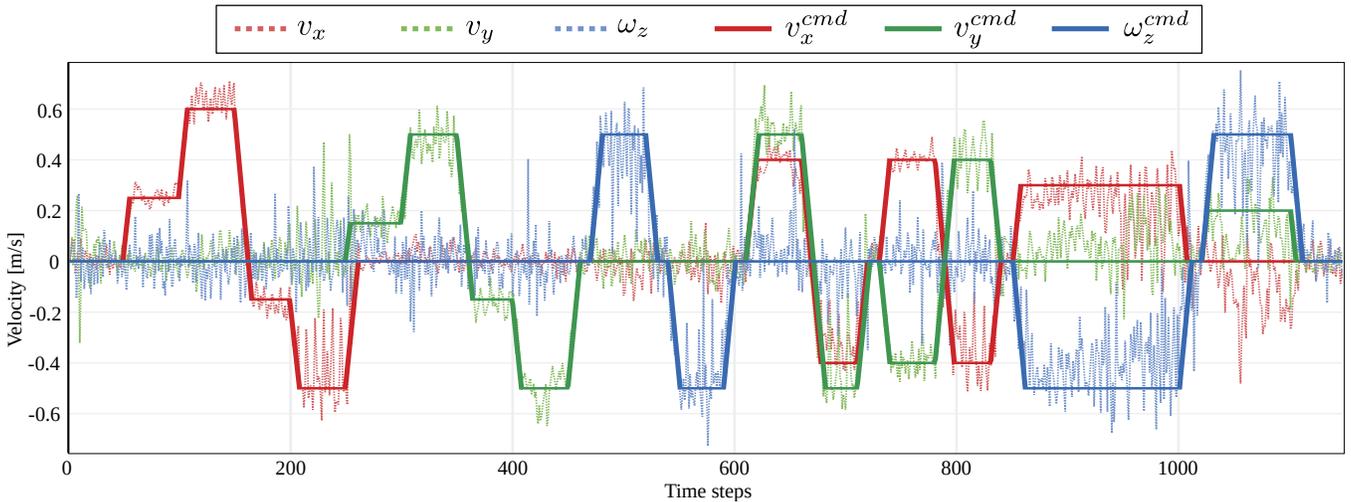


Fig. 3. Velocity tracking. The learned control policy is able to follow the commanded velocities (solid wider lines) in different directions with varying velocities. Initially, motions in single directions (sagittal, lateral and turning) are demonstrated. After 600 time steps, motions with combined directions (sagittal with lateral, sagittal with turning and lateral with turning) are evaluated.

against actuation latency, a latency value  $t_{lat}$  is uniformly sampled  $t_{lat} \sim \mathcal{U}(t_{lat}^{min}, t_{lat}^{max})$  at the beginning of each training episode.

5) *Network Output Filtering*: We filter the actions inferred by the policy before sending the corresponding commands to the actuators using a Butterworth low-pass filter.

## V. EVALUATION

Our approach is evaluated on the NimbRo-OP2X robot [21]. The robot is 135 cm high and weighs 19 kg. The platform possesses 18 Degrees of Freedom (DoF): five on each leg, three on each arm and two for the head. The joints at the head (pitch and yaw) are neglected because their contribution for walking is considered insignificant and they are needed for active visual perception. Each leg exhibits a parallel kinematic structure containing: a hip yaw, a hip roll, a hip pitch, a knee pitch and an ankle roll. For state estimation, the robot uses only a gyroscope and an accelerometer enabling a 2 DoF state estimation, namely the roll and the pitch angles. The action space was bounded to  $\delta \in [-0.1, 0.1]$ . In total, our robot creates a 47-dimensional state space and a 16-dimensional action space.

The training is carried out in the physics-based simulator MuJoCo. The simulation runs at 1 kHz. The PD controllers have the same frequency as the simulator. The task was implemented as an OpenAI Gym environment [22]. To speed up the training, 12 parallel environments were running simultaneously on an Intel i9-9900K CPU. The weights of the reward function terms ( $w_v = 42, w_r = 4, w_a = 4$  and  $w_f = 18$ ) were set manually from experience. The kernel sensitivities were set to  $l_v = 9, l_r = 3$  and  $l_f = 10$ .

The networks are trained iteratively by epochs. Each epoch comprises 800 time steps. Both networks use a learning rate equal to  $1 \times 10^{-4}$  with a batch size of 480. Per epoch, 10 updates are performed using the Adam optimizer. Finally, the decaying factor  $\gamma$  is set to 0.99 and  $\tau = 0.97$ . The training finished after 7400 epochs for a total of  $7.1 \times 10^7$  time steps,

resulting in 20.5 days of simulated time which corresponds to 32.5 hours of computation in real time.

The robot is asked to learn to walk at  $v_{core} = [0.4 \text{ m/s}, 0.0 \text{ m/s}, 0.0 \text{ rad/s}]$ . The minimum and maximum velocities in each direction are bounded to  $v_x \in [-0.6, 0.6]$ ,  $v_y \in [-0.6, 0.6]$ , and  $\omega_z \in [-0.6, 0.6]$ . Finally, the curriculum variables of the reward terms are initialized at  $C_v = 0.01$  with  $k_d = 0.95$ , and  $C_{foot} = 0.05$  with  $k_d = 0.995$ .

The factor noise to the PD controllers  $\epsilon_{pd}$  is set to 0.1. The standard deviations of the noise applied to the sensory data are:  $\eta_g = 1 \times 10^{-4}$ ,  $\eta_a = 1 \times 10^{-4}$ ,  $\eta_q = 1 \times 10^{-3}$ , and  $\eta_v = 1 \times 10^{-3}$ . The friction values are sampled from  $\mu_t \sim \mathcal{U}(0.4, 0.8)$  and  $\mu_z \sim \mathcal{U}(0.1, 0.3)$ , for the tangential, and torsional friction coefficients. The injected noise and sampled friction values are considered only in simulation. The latency  $t_{lat}$  is sampled uniformly from  $[0, 50]$ ms in training ( $t_{lat} = 8$ ms for the real robot). Finally, the cutoff frequency of the low-pass filter is set to 10 Hz for the real robot.

Initially, we evaluate the learned gait in simulation. Figure 3 shows the velocity of the base given commanded

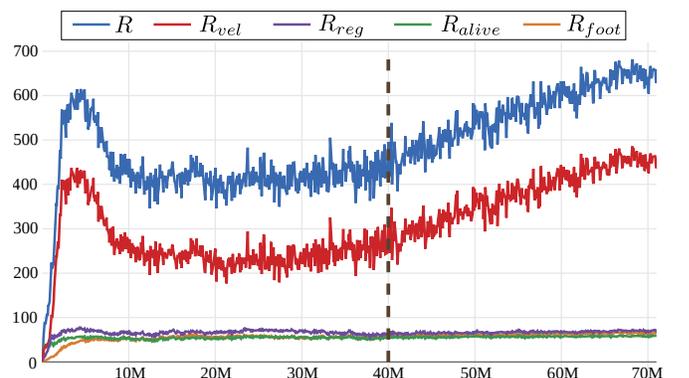


Fig. 4. Training return curves. The vertical dashed line represents the point where the limits of the velocity scheduler have been reached. Note how the agent continues refining the policy after these limits have been reached.

TABLE I  
RESULTS OF FALLING TEST

Commanded Velocity	# Falls	Commanded Velocity	# Falls
$v = [0.0, 0.0, 0.0]$	0/10	$v = [0.0, 0.0, 0.6]$	1/10
$v = [0.6, 0.0, 0.0]$	0/10	$v = [0.0, 0.0, -0.6]$	3/10
$v = [-0.6, 0.0, 0.0]$	2/10	$v = [0.4, 0.4, 0.0]$	0/10
$v = [0.0, 0.6, 0.0]$	0/10	$v = [0.4, 0.3, 0.0]$	0/10
$v = [0.0, -0.6, 0.0]$	0/10	$v = [0.0, 0.4, 0.3]$	0/10

velocities. Figure 4 shows the return training curves. At the end of training, the robot falls rarely having an alive reward value oscillating around its maximum (80). Falls happen mostly when robot is commanded to the velocity limits. Moreover, the robot learned to walk at different speeds in different directions. Interestingly, the robot also learns to walk in place, i.e., to lift the feet rhythmically without moving in any direction. The accompanying video shows the acquired locomotion skills.

In order to evaluate the robustness of the learned controller, we performed ten walking sequences of 60s for different commanded velocities and counted the number of falls. The results are presented in Table I. The robot was more susceptible to falling when going backwards and turning along the vertical axis. The agent had less experiences walking backwards compared to the forward counterpart. Regarding turning, the robot was requested to move at the target velocity limits which suggests an increment in these limits. Additionally, the capacity of the controller to switch between velocities was evaluated by sampling uniformly 3-dimensional commanded velocities from the limits defined above. From 160 changes, the robot successfully performed 150 establishing a 93.75% success rate.

In addition, our learned controller was evaluated against perturbations. The robot was pushed at the base link frame for 0.2s from the front, the back and the side at different commanded velocities. For each commanded velocity, we started perturbing the robot with 10N pushes and increased this magnitude by 10N after ten pushes. Table II presents the maximum push the robot was able to reject successfully 10 times in a row (100% Succ.), and the maximum push the robot was able to reject at least once (Max. push).

In order to evaluate the contribution of the velocity scheduler, we trained a policy without curriculum. After the same number of epochs, the robot learned to stand but it was not able to walk in any direction. Additionally, a controller was learned replacing the Beta policy by a Gaussian one. With this controller, the robot was able to stand but it was not able to go more than four steps without falling due to constant saturation of the PD controllers. This demonstrates that the introduction of Beta policies plays a key role on the use of energy and furthermore it avoids the incorporation of torque terms in the reward function and the corresponding weight assignment and torque measurements or estimates.

Finally, we transfer the learned gait to real hardware. The NimRo-OP2X robot is able to walk on the spot, forwards, backwards, laterally and diagonally. Turning in place is also possible. We observed dissimilarities between the simulated and the real gait which demands for more sophisticated sim-

TABLE II  
RESULTS OF PERTURBATION [N] TEST

Commanded velocity	Test	Front push		Lateral p.	
				right	left
In-place	100% Succ.	40	20	50	
	Max. push	60	30	90	
$v = [0.3, 0, 0]$	100% Succ.	30	20	50	
	Max. push	50	30	80	
$v = [0.6, 0, 0]$	100% Succ.	30	10	40	
	Max. push	50	20	70	
$v = [0, 0.25, 0]$	100% Succ.	40	20	40	40
	Max. push	60	30	60	100
$v = [0, 0.5, 0]$	100% Succ.	40	20	20	50
	Max. push	60	30	60	100

to-real transfer methods. Snapshots of the robot walking forward and lateral are shown in Fig. 5 and Fig. 6.

## VI. CONCLUSION

We presented a novel approach to learn a single control policy capable of omnidirectional walking for humanoid robots using a realistic robot model. We have demonstrated the capacity of the learned policy to walk in the sagittal and lateral directions and to turn around the vertical axis at different speeds. Without altering the policy, our approach also produces motions in combined directions, i.e., the agent is able to walk diagonally and to turn during walking. Achieving these locomotion behaviors was possible mainly due to: the velocity scheduler, the introduction of a core velocity; the use of beta policies to bound the action space; the incorporation into the state  $s_t$  of the velocity of the base expressed in the relative and in the reference frame. Our approach does not require reference motions to achieve anthropomorphic locomotion thanks to the introduction of a nominal pose.

In the future, we will explore different alternatives for motion regularizers to find one that is more flexible than motion capture data and generates more anthropomorphic motions compared to nominal poses. In addition, we would like to extend this approach for 3D walking, such that the robot is able to walk on sloped terrains and to climb stairs.



Fig. 5. Snapshots of forward walk performed by the real robot commanded by the learned locomotion controller.



Fig. 6. Snapshots of lateral (left) walk performed by the real robot commanded by the learned locomotion controller.

## REFERENCES

- [1] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, "Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.
- [2] Z. Xie, G. Berseth, P. Clary, J. Hurst, and M. van de Panne, "Feedback control for Cassie with deep reinforcement learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.
- [3] Z. Xie, P. Clary, J. Dao, P. Morais, J. Hurst, and M. van de Panne, "Learning locomotion skills for Cassie: Iterative design and sim-to-real," in *Conference on Robot Learning (CORL)*, 2019.
- [4] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, "Sim-to-real: Learning agile locomotion for quadruped robots," in *Robotics: Science and Systems (RSS)*, 2018.
- [5] P.-W. Chou, D. Maturana, and S. Scherer, "Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution," in *34th International Conference on Machine Learning (ICML)*, 2017.
- [6] D. Rodriguez, A. Brandenburger, and S. Behnke, "Combining simulations and real-robot experiments for Bayesian optimization of bipedal gait stabilization," in *Proceedings of RoboCup International Symposium*, 2018.
- [7] A. Rai, R. Antonova, S. Song, W. C. Martin, H. Geyer, and C. G. Atkeson, "Bayesian optimization using domain knowledge on the ATRIAS biped," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018.
- [8] H. Farazi, F. Ahmadinejad, F. Maleki, and M. Shiri, "Evolutionary approach for developing fast and stable offline humanoid walk," in *3rd Joint Conference of AI & Robotics and 5th RoboCup Iran Open International Symposium*, 2013.
- [9] N. Heess, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami, M. Riedmiller, et al., "Emergence of locomotion behaviours in rich environments," *CoRR abs/1707.02286*, 2017.
- [10] Zhaoming, H. Y. Ling, N. H. Kim, and M. van de Panne, "ALLSTEPS: Curriculum-driven Learning of Stepping Stone Skills," in *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2020.
- [11] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.
- [12] K. Bergamin, S. Clavet, D. Holden, and J. R. Forbes, "DReCon: data-driven responsive control of physics-based characters," *ACM Transactions On Graphics (TOG)*, vol. 38, no. 6, pp. 1–11, 2019.
- [13] S. Park, H. Ryu, S. Lee, S. Lee, and J. Lee, "Learning predict-and-simulate policies from unorganized human motion data," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–11, 2019.
- [14] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, 2019.
- [15] C. Yang, K. Yuan, W. Merkt, T. Komura, S. Vijayakumar, and Z. Li, "Learning whole-body motor skills for humanoids," in *IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018.
- [16] V. Tsounis, M. Alge, J. Lee, F. Farshidian, and M. Hutter, "DeepGait: Planning and control of quadrupedal gaits using deep reinforcement learning," *IEEE Robotics and Automation Letters*, 2020.
- [17] W. Yu, V. C. Kumar, G. Turk, and C. K. Liu, "Sim-to-real transfer for biped locomotion," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.
- [18] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *International Conference on Learning Representations (ICLR)*, 2016.
- [19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR abs/1707.06347*, 2017.
- [20] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *26th annual International Conference on Machine Learning (ICML)*, 2009.
- [21] G. Ficht, H. Farazi, A. Brandenburger, D. Rodriguez, D. Pavlichenko, P. Allgeuer, M. Hosseini, and S. Behnke, "NimbRo-OP2X: Adult-sized open-source 3D printed humanoid robot," *IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018.
- [22] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, *OpenAI Gym*, 2016. eprint: arXiv:1606.01540.