Learning to Walk With Deep Reinforcement Learning: Forward Dynamic Simulation of a Physics-Based Musculoskeletal Model of an Osseointegrated Transfemoral Amputee

Brown N. Ogum, Lambert R. B. Schomaker, *Senior Member, IEEE*,
and Raffaella Carloni, *Member, IEEE*

***Abstract*—This paper leverages the OpenSim physics-based simulation environment for the forward dynamic simulation of an osseointegrated transfemoral amputee musculoskeletal model, wearing a generic prosthesis. A deep reinforcement learning architecture, which combines the proximal policy optimization algorithm with imitation learning, is designed to enable the model to walk by using three different observation states. The first is a complete state that includes the agent's kinematics, ground reaction forces, and muscle data; the second is a reduced state that only includes the kinematics and ground reaction forces; the third is an augmented state that combines the kinematics and ground reaction forces with a prediction of the muscle data generated by a fully-connected feed-forward neural network. The empirical results demonstrate that the model trained with the augmented observation state can achieve walking patterns with rewards and gait symmetry ratings comparable to those of the model trained with the complete observation state, while there are no symmetric walking patterns when using the reduced observation state. This paper shows the importance of including muscle data in a deep reinforcement learning architecture for the forward dynamic simulation of musculoskeletal models of transfemoral amputees.**

***Index Terms*—Deep reinforcement learning, computer simulation, prosthetics.**

I. Introduction

COMPUTER simulations, of either finite element models or musculoskeletal models, provide a valuable tool for studying how impaired gaits manifest among individuals with transfemoral (above-knee) amputations [1], [2], as well as for understanding how prosthetic devices can affect walking patterns and enhance users' mobility [3]. More specifically, inverse dynamic simulations have proven instrumental in analyzing the biomechanics of the gait of transfemoral amputees [4], diagnosis [5], rehabilitation [6]. Conversely, forward dynamic simulations represent a more complex task, as they require to modulate the activation or deactivation of specific muscle groups in order to generate desired patterns throughout the gait cycle in close cooperation with the control of a prosthesis [7].

Inspired by the idea that humans learn to walk by interacting with the surrounding environment, Deep Reinforcement Learning (DRL) has the potential of being successfully used for performing forward dynamic simulations of physics-based musculoskeletal models [8], [9]. The knowledge developed by a human model (i.e., the simulated agent) from experiencing the surrounding environment (which includes a prosthetic device, in case of individuals with a transfemoral amputation) helps to find a policy that, while maximizing a predefined reward, computes the muscles' activation/deactivation and the forces/torques at the prosthetic joints. As a result, the learned policy will make the agent walk [8], [10].

In the current literature, DRL has showcased progress in solving simulated control problems for simple humanoids and bipedal robots [11], [12], [13], [14]. However, the utilization of DRL in simulations of physics-based musculoskeletal models has remained limited. In [15], a hierarchical structure of policy networks is introduced, where the skeletal part learns the kinematics and dynamics through a Markov decision process, while, subsequently, the muscular part learns the muscle activations through quadratic programming. In our previous work, a DRL architecture was proposed that combines Proximal Policy Optimization (PPO) [16] with imitation learning [17] for normal walking of healthy and impaired human musculoskeletal models [8], and for ramp/stair ascending of a healthy musculoskeletal model [9].

Building upon our previous work, this study proposes a novel DRL architecture to teach a transfemoral amputee musculoskeletal model, wearing a generic prosthesis, to walk. As shown in Figure 1, the agent in the open-source simulation environment OpenSim (NIH National Center for Simulation in Rehabilitation Research, Stanford, CA, USA, www.opensim.stanford.edu) is trained by a deep neural

Manuscript received 12 July 2023; revised 13 December 2023; accepted 5 January 2024. Date of publication 10 January 2024; date of current version 24 January 2024. This work was supported by the European Commission's Horizon 2020 Program as part of the Project MyLeg under Grant 780871. *(Corresponding author: Raffaella Carloni.)*

The authors are with the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Faculty of Science and Engineering, University of Groningen, 9747 AG Groningen, The Netherlands (e-mail: brownyogum@gmail.com; l.r.b.schomaker@rug.nl; r.carloni@rug.nl).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2024.3352416, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2024.3352416

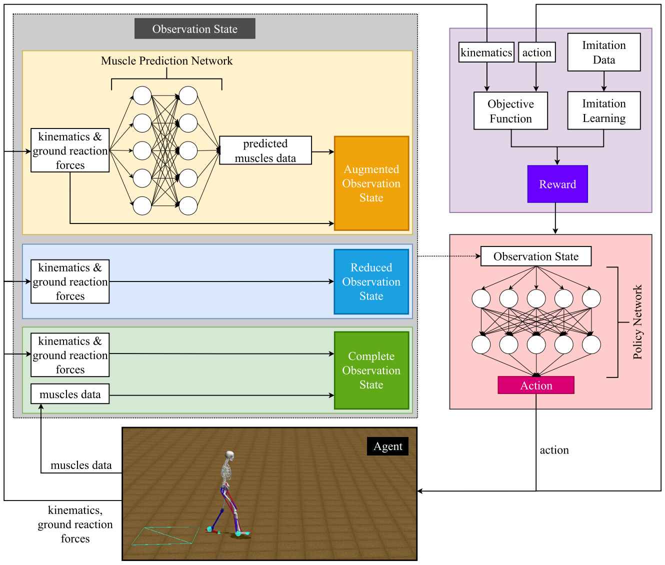

Observation State
Muscle Prediction Network
kinematics & ground reaction forces
predicted muscles data
Augmented Observation State
kinematics & ground reaction forces
Reduced Observation State
kinematics & ground reaction forces
muscles data
Complete Observation State
kinematics
action
Imitation Data
Objective Function
Imitation Learning
Reward
Observation State
Policy Network
Action
action
Agent
muscles data
kinematics, ground reaction forces

Fig. 1. The DRL architecture with the three different implementations of the observation state: complete (kinematics, ground reaction forces, and muscle data), reduced (kinematics and ground reaction forces), and augmented observation state (kinematics, ground reaction forces, and a prediction of the muscle data by means of a neural network). The reward of the DRL architecture is computed on an objective function and on imitation learning.

TABLE I
DEGREES OF FREEDOM AND RANGE OF MOTION OF THE OSSEOINTEGRATED TRANSFEMORAL AMPUTEE MODEL `gait1415+2`

Body	DOFs	Joint [Range of motion]
Pelvis	6	pelvis_tilt, pelvis_list, pelvis_rotation [$-90°$, $90°$] pelvis_x, pelvis_y, pelvis_z
Hip	2+2	hip_flexion [$-120°$, $120°$], hip_adduction [$-45°$, $45°$]
Knee	1+1	knee_angle (flexion/extension) [$-120°$, $10°$]
Ankle	1+1	ankle_angle (dorsiflexion/plantarflexion) [$-60°$, $30°$]

TABLE II
PRIMARY FUNCTIONS OF THE 15 MUSCLES AND 2 ACTUATORS OF THE MODEL

Muscles & Actuators	Primary Functions	Leg
Gluteus maximus	Hip extension	both
Iliopsoas	Hip flexion	both
Hip abductor	Hip abduction	both
Hip adductor	Hip adduction	both
Hamstring (biarticular)	Hip extension, knee flexion	right
Rectus femuris (biarticular)	Hip flexion, knee extension	right
Vasti	Knee extension	right
Biceps femoris	Knee flexion	right
Soleus	Ankle extension (plantarflexion)	right
Gastrocnemius (biarticular)	Knee flexion, ankle extension	right
Tibialis anterior	Ankle flexion (dorsiflexion)	right
Knee actuator	Knee flexion/extension	left
Ankle actuator	Ankle dorsiflexion/plantarflexion	left

network that receives the agent's observation state and computes an action (i.e., the muscle forces and the torques of the actuators of the prosthesis). The deep neural network also receives, as an input, a reward which is computed on the agent's kinematics and the agent's action according to an objective function and imitation data. Three distinct implementations of the DRL architecture are developed and compared. The key differentiating factor among these implementations lies in the utilization of three different observation states for the musculoskeletal model. The first is a *complete state* that includes the agent's kinematics, ground reaction forces (GRFs), and muscle data (force, length, velocity), as done in the current literature [8], [9], [15]. The second is a *reduced state* that only includes the kinematics and GRFs; this choice was made to exclude muscle data, which can be difficult to measure and process in real scenarios, but that can be derived from the simulation [18]. The third is an *augmented state* that includes the kinematics, GRFs, and a prediction of the muscle data, generated by a feed-forward neural network that uses only the kinematics and the GRFs.

Empirical results show that the transfemoral amputee musculoskeletal model, trained with the proposed DRL architecture with the augmented observation state, achieves walking patterns with rewards and gait symmetry ratings comparable to those of the model trained with the complete observation state, while there are no symmetric walking patterns when using the reduced observation state. These results highlight the importance of integrating muscle data into the DRL architecture and, in scenarios where such data are not readily accessible via sensors, leveraging artificial intelligence methods for predicting muscle data becomes crucial. This finding represents a significant leap forward in the realm of eventually exploiting physics-based simulations of musculoskeletal models and deep reinforcement learning for the control of prosthetic limbs in individuals with transfemoral amputations.

II. THE MUSCULOSKELETAL MODEL

The transfemoral amputee model used in this study has been derived from the OpenSim 4.2 lower-extremity musculoskeletal model `gait1422` of a healthy subject. Specifically, seven musculotendon units were removed from the amputated (left) leg, and a generic bone-anchored transfemoral prosthesis (with one ideal actuator at the knee joint and one at the ankle joint) was added. This model, named `gait1415+2`, is a simplified abstraction of the model presented in our previous work [2], and has been developed to perform forward dynamic simulations by means of computationally-demanding artificial intelligence methods [19].

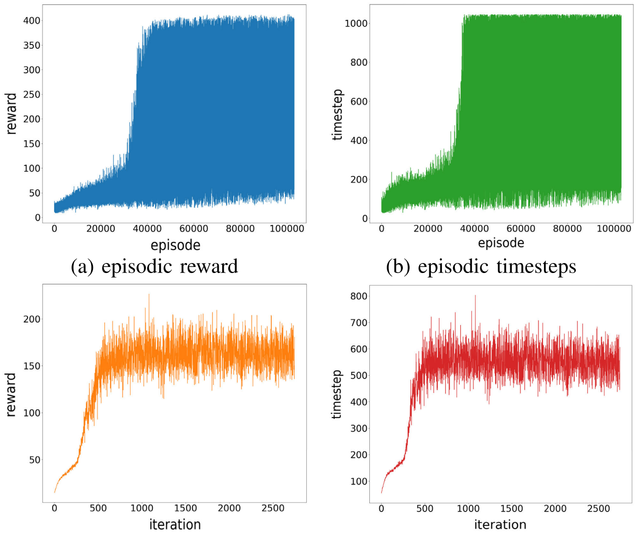
The model has 14 degrees of freedom (DOFs), 15 musculotendon units, and 2 actuators. Table I summarizes the 14 DOFs (6 for the pelvis, 2 for each hip joint, 1 for each knee joint, 1 for each ankle joint), and their range of motion. The lumbar extension is locked to $-5°$, while the hip rotation is locked to $0°$ (as in the `gait1422` model). Table II summarizes the 15 musculotendon units and the 2 actuators at the knee and ankle joints, together with their primary function. The musculotendon units are modeled as Hill-type muscles with a non-linear first-order dynamics between excitation and activation [20]. The muscle activations, which can range between 0% and 100%, generate a muscle force as a function of the muscle physical properties. The 2 ideal actuators are modelled as OpenSim activation coordinate actuators, which produce a generalized force with a first-order linear activation dynamics.

A. The Imitation Data-Set

The imitation data-set used in this study is an experimental public data-set that has been collected on 83 typically developing children by measuring the three-dimensional lower extremity joint kinematics, joint kinetics, surface electromyographic, and spatio-temporal data [21]. The data-set contains the means and the ±1 standard deviations of all subjects over one gait cycle at speeds ranging from very slow to slow, fast, and very fast with respect to the free speed. It should be noted that this data-set required no scaling because the mean values match the dimensions of the `gait1415+2` model. However, the data were processed to guarantee a symmetric gait pattern for the hip, knee, and ankle joints [8].

III. Deep Reinforcement Learning Architecture

This Section details the DRL architecture (Figure 1) that is proposed in this study to teach the agent (osseointegrated transfemoral amputee model) to perform a human-like gait. It is important to note that the proposed DRL method is general and could be applied, with the necessary adaptations, to a different physics-based musculoskeletal model.



(a) episodic reward

(b) episodic timesteps

(c) average episodic reward per iteration

(d) average episodic timesteps per iteration

Fig. 2. Rewards and timesteps per episode and iteration. (top) Rewards and duration per episode. (bottom) Average episodic reward and average episode duration during each training iteration.

A. Proximal Policy Optimization

The DRL architecture relies on a feed-forward artificial neural network, hereafter called *policy network*, with four layers: one input layer (i.e., the agents' observed state), two hidden layers, and an output layer (i.e., the agent's action) [8].

To teach the agent to walk, the policy network is trained with PPO [16]. Let $r_t(\theta)$ be the ratio between the probabilities of the new and old policy, i.e.:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

where θ and θ_{old} are the new and the old parameters, π_θ is the policy corresponding to the parameters θ, a_t and s_t the action and the state vectors at the time-step t, respectively. PPO uses the following objective function:

$$L^{CLIP}(\theta) = \mathbb{E}\left[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

where $\mathbb{E}$ is the expected value, $\hat{A}_t$ is the advantage estimation, i.e., the difference between the expected and the real reward from an action, and ϵ is the clip value. If the probability ratio falls outside the range $[(1-\epsilon), \cdots, (1+\epsilon)]$, the advantage function is clipped to prevent too large policy updates.

1) Hyperparameter - Iterations: The first hyperparameter to select in the course of training the neural network is the amount of learning iterations to perform. In this study, a PPO training iteration is characterized by the optimization performed on 15360 simulation timesteps (i.e., the simulated agent's state at 10 ms intervals), using a batch size of 512 [8].

Figure 2 shows the episodic reward and episodic timesteps during the training of the policy network, as well as the average episodic reward and average episodic timesteps per iteration. From Figure 2c, it can be observed that there is a steep increase in the average episodic reward obtained in the first ∼500 iterations of the training. This is also reflected in the steady increase in the total episodic rewards obtained in the first ∼39000 episodes of the simulation (Figure 2a) and in the number of timesteps (Figures 2b and 2d). After ∼39000 episodes, there is no improvement in the total timesteps. This is expected as the maximum number of timesteps is bounded by the size of the imitation data (1050). In the episodic reward and episodic timesteps (Figures 2a and 2b), after the steep learning phase, there is a noticeable variance, which is expected due to the random initialization of the agent during each episode of training. Reward values ranging from ∼20 to 400 can be observed as well as timesteps values from ∼20 to 1050. This is due to the exploration by the learning algorithm even after finding a solution that completes the learning task. The stochastic exploratory actions try to find an even better solution but, as it can be observed in Figure 2c, a better solution cannot be reliably found as there is no increasing trend in the learning curve for the realized reward per iteration even after 2500 iterations. On the contrary, there is a slightly decreasing trend in both the average rewards and timesteps per iteration after the first 700 iterations. This is likely due to the learning algorithm trying to further maximize its immediate rewards at a detriment to the robustness, hence showing early signs of overfitting to the imitation data. Therefore, 700 is chosen as the number of iterations to perform the training of the agent, namely the number of iterations for which a high average episodic reward and timesteps have been obtained.

2) Hyperparameter - Hidden Units: Different policy networks were trained using different numbers of hidden units, as shown in Table III. The policy network was trained for 700 iterations, and the rewards for 50 episodes were averaged. The policy network with 228 hidden units resulted in the highest average reward, and chosen for the final DRL architecture [8].

3) Hyperparameter - Prediction Categories: Table IV shows the average reward per episode for different numbers of

TABLE III

Average Episode Reward and Training Duration for the Different Number of Hidden Units in the Policy Network. The Episode Rewards Are Averaged Over 50 Episodes After 700 Iterations of PPO

N. hidden units	Average episode rewards	Training duration (h)
100	156.2	51.9
228	163.5	52.2
312	158.3	53.3

TABLE IV

Average Episode Reward and Training Duration for the Different Number of Prediction Categories for Training With PPO. The Episode Rewards Are Averaged Over 50 Episodes After 700 Iterations of PPO

N. prediction categories	Average episode rewards	Training duration (h)
2	163.5	52.2
3	194.3	52.4
5	198.7	54.9
7	187.0	55.3
9	169.9	56.8
15	153.1	58.9
21	162.8	63.9

TABLE V

Hyperparameters Used for the Training of the DRL Architecture

Parameter	Value
Workers	10
Iterations	≥ 700
Value network optimizer	Stochastic grad. desc.
Policy network optimizer	Adam
Steps per worker	1,536
Steps per action	1
Entropy coefficient	0.01
Stochastic policy	true
Prediction categories	≥ 2
PPO clip parameter (ϵ)	0.2
Optimization mini-batch size	512
N. of policy updates per mini-batch	4
Discount factor	0.99
Generalized advantage estimate	0.95
Hidden layers in policy and value networks	2
Hidden layer size	[100, 228, 312]
Activation function	tanh
Optimization stepsize	0.001

prediction categories, as discretization of the action space. The policy network was trained for 700 iterations, and the rewards for 50 episodes were averaged. A policy network with 5 prediction categories had the best performance, and chosen in the final DRL architecture [8].

Table V summarizes the chosen final hyperparameters of the PPO implementation.

B. Observation State

The policy network is trained with three different observation states, which means that the input layer of the neural network has three different dimensions, as detailed hereafter.

TABLE VI

Combination of Hyperparameters Used to Design the Muscle Prediction Network

Parameter	Value(s)
Optimization algorithm	Adam
Loss function	Mean absolute error
Number of hidden layers	1, 2, 3
Number of hidden layer units	64, 256, 512
Activation function	relu, tanh
Batch normalization	with, without
Drop out	with, without
Early Stopping	yes (20 epochs)

1) Complete Observation State: The first observation state is a *complete state*, which includes the agent's kinematics, GRFs, and muscle data [8]. In this case the dimension of the input layer of the policy network is 91. The 46 states for the kinematics and GRFs are: 6 positions and 6 velocities of the pelvis, 4 positions and 4 velocities of the hips, 2 positions and 2 velocities of the knees, 2 positions and 2 velocities of the ankles, 6 GRFs for the feet, 6 actuation data (force, velocity, control, power, activation, actuation from the OpenSim activation coordinate actuator), of the knee and 6 for the ankle. The 45 states for the muscle data are the 3 data (fiber force, fiber length, fiber velocity) for each one of the agent's 15 muscles.

2) Reduced Observation State: The second observation state is a *reduced state*, which includes only the agent's kinematics and the GRFs. In this case the dimension of the input layer of the policy network is 46, which has decreased by removing the muscle data.

3) Augmented Observation State: The third observation state is an *augmented state*, which includes the agent's kinematics, the GRFs, and a prediction of the muscle data. In this case the dimension of the input layer of the policy network is 91, as for the complete observation.

To predict the muscle data, the reduced observation state is fed to a pre-trained fully-connected feed-forward artificial neural network, called *muscle prediction network*. Table VI summarizes the hyperparameters that have been compared to select the best performing network. Two million observation steps have been collected using random initialization and action sampling for the agent in the overall DRL architecture. The collected two million time-steps of data have been split into 90% training and 10% testing [10]. The hyperparameter selection was done on the training split by using every combination of hyperparameter values in Table VI. The best model on cross-validation is then tested on the remaining 10% of the data split to give a non-biased performance estimate. By using a 5-fold cross-validation loss (mean and absolute error), it was observed that a fully-connected network with 3 hidden layers, 256 hidden units, batch normalization, drop-out, and ReLU activation had the best performance with a mean absolute error of 0.11. The final network, chosen for the model data prediction, is shown in Figure 3. This network has an average mean absolute error of 0.082 N on the prediction of the normalized 15 muscle forces, a 0.018 m mean absolute error for the prediction of the normalized length of the muscles, and a 0.228 m/s mean absolute error for the prediction of the muscles' velocities.

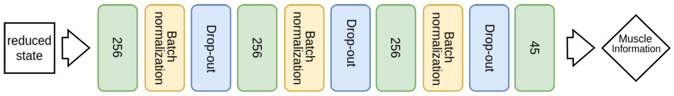


Fig. 3. Muscle prediction network. The green blocks are the fully-connected layers (indicating the number of units); the yellow blocks are the batch normalization; the blue blocks are the drop-out regularization.

C. Action State

The dimension of the output layer of the policy network is 17, which corresponds to the agent's 15 muscles and the 2 actuators of the prosthesis. The action performed by the agent is bounded by laws of physics in OpenSim.

D. Reward

This study uses the reward function $r_t = 0.1 \cdot r_{goal,t} + 0.9 \cdot r_{imitation,t} - p_t$, which consists of a goal reward $r_{goal,t}$, an imitation reward $r_{imitation,t}$, and a penalty term p_t. At each timestep t, the instantaneous reward r_t is computed for a given state s_t, action a_t, and consequent state s_{t+1}. The weights were chosen so the agent learns a human-like gait quicker.

1) Goal Reward: The goal reward at timestep t creates an incentive for the agent to move in a continuous straight direction. The goal reward $r_{goal,t} = e^{-(p_x+p_y+p_z)}$ is obtained by computing the error p between the actual and desired values of the pelvis' coordinates pelvis_x, pelvis_y, pelvis_z.

2) Imitation Reward: The imitation reward $r_{imitation,t}$ at timestep t is issued to create an incentive for the agent to mimic the kinematics provided by the imitation data. The imitation reward is given by $r_{imitation,t} = 0.9 \cdot r_{position,t} + 0.1 \cdot r_{velocity,t}$. The imitation position reward $r_{position,t}$ and the imitation velocity reward $r_{velocity,t}$ are obtained by comparing the position and velocity of the different joints of the agent (hip_flexion, hip_adduction, knee_angle, ankle_angle of both legs) at timestep t with that of the imitation data. The weights were chosen so the agent learns a human-like gait quicker.

3) Penalty: The penalty p_t at timestep t encourages the agent to find an energy-efficient manner of solving the optimization task by penalizing the action (the agent's 11 muscles and prosthesis' 2 actuators) at each timestep, thereby, limiting the reward obtained for actions with high activation values.

E. Performance Criteria

The performance of the agent is evaluated with respect to the average episodic reward of the DRL architecture over 50 episodes of simulations, and with respect to the symmetry of the gait over 50 gait cycles, which is analyzed by using three different metrics as explained hereafter.

1) RMSE: It is given by $\sqrt{\sum_{t=1}^{T}(x_{\mathrm{ob},t} - x_{\mathrm{im},t})^2/T}$, where $x_{\mathrm{ob},t}$ and $x_{\mathrm{im},t}$ are the observed state and the imitation data at time t, respectively. This metric has the advantage of being in the same unit as the observed state, but it is not normalized.

2) Symmetry Angle: It quantifies the gait symmetry [22], and is given by $[45° - atan(X_a/X_u)]/90° \cdot 100\%$, where X_a and X_u are the two angles to compare. A symmetry angle of 0% indicates perfect symmetry, while 100% indicates that the angles are equal but opposite in magnitude.

TABLE VII
AVERAGE EPISODIC REWARD FOR THE DIFFERENT NUMBER OF PREDICTION CATEGORIES AND OBSERVATION SPACE TYPES. THE REWARDS ARE AVERAGED OVER 50 EPISODES AFTER 700 ITERATIONS OF PPO

N. prediction categories	**Average episode reward**		
	complete observation	reduced observation	augmented observation
2	163.5	142.1	166.7
3	194.3	161.2	185.2
5	198.7	160.5	190.9

3) Trend Symmetry: It is a metric to evaluate the joint angle symmetry using two time series of joint angle data [23]. The trend symmetry ranges from 0 to 1, where 0 indicates a perfect symmetry. This metric is important in this study because the RMSE and symmetry angle neglect the temporal information in the gait waveforms. To compute the trend symmetry, eigenvectors are used to compare time-normalized gait cycle data. The trend symmetry uses the principal eigenvector to analyze the variance of the distribution of the points formed by the pair of waveforms to be compared, and it provides a measure of symmetry that is not affected by the difference in magnitude between the two waveforms. In addition, the trend symmetry computes the symmetry of two waveforms using the entire waveforms.

IV. RESULTS

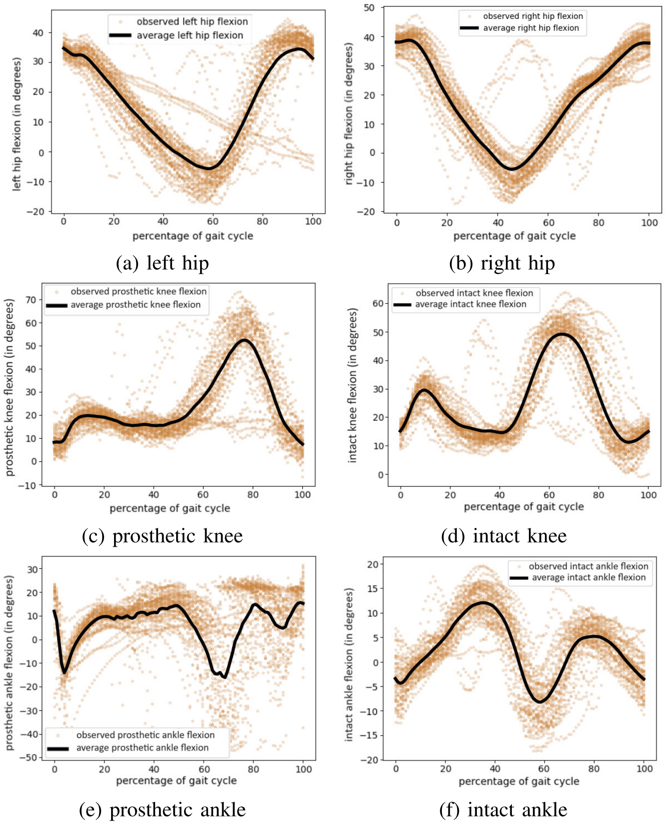
A. Average Episodic Reward With Different Observation States

This Section compares the emerging rewards when using the complete, reduced, and augmented observation spaces. The DRL architecture has been trained for 700 iterations of PPO, using different numbers of prediction categories, and the average reward over 50 episodes is computed, as reported in Table VII. The most average episodic reward is obtained with the complete observation state with 5 prediction categories (198.7), while with the augmented observation state it is 190.9.

B. Forward Dynamics

This Section reports the results of the forward dynamics simulation (hip, knee, and ankle angles) during the gait cycle, when the three different observations states are used.

1) Complete Observation: Figure 4 shows the instantaneous and average hip, knee, and ankle joints angles for both legs of the agent during 50 gait cycles when a complete observation state is used. Figures 4a and 4b show high symmetry between the hips. There is a little variance in the emerging pattern, meaning that the agent exhibits a natural walking pattern without much deviation during the gait. Similarly for the knees in Figures 4c and 4d, a few deviations are observed with a peak at about 20% and 40% of the gait cycle. Moreover, it can be observed that there is a very high symmetry between the intact and the prosthetic knee, resulting in a natural walking pattern. There is, however, much less symmetry between the intact and the prosthetic ankle; while the intact ankle generates a healthy ankle angle during the gait cycle, that is not the

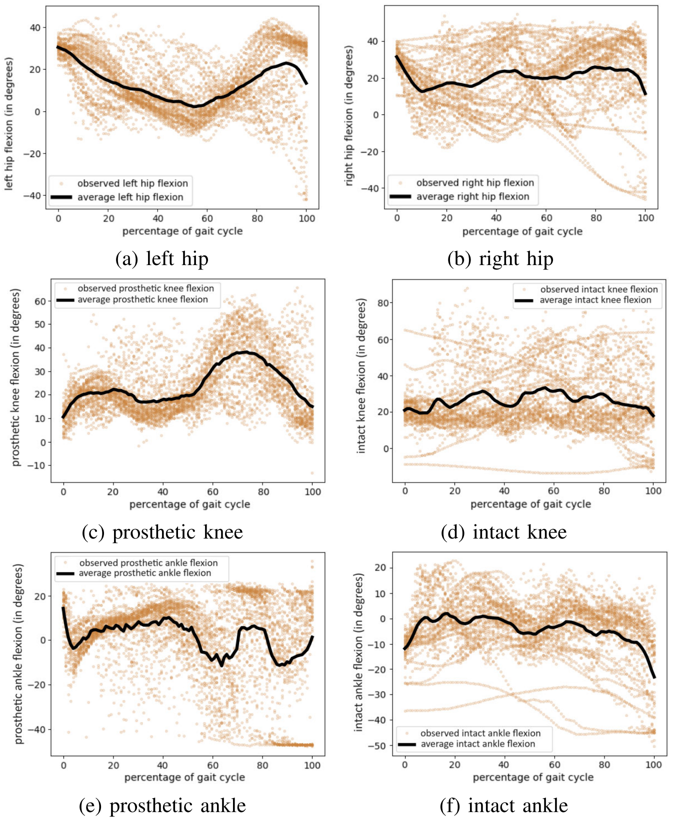


(a) left hip (b) right hip

(c) prosthetic knee (d) intact knee

(e) prosthetic ankle (f) intact ankle

Fig. 4. Angles of the hips, knees, and ankles during the gait cycle using the complete state observation. The policy network has 228 hidden units and 5 prediction categories, and was trained for 700 iterations of PPO.



(a) left hip (b) right hip

(c) prosthetic knee (d) intact knee

(e) prosthetic ankle (f) intact ankle

Fig. 5. Angles of the hips, knees, and ankles during the gait cycle using the reduced state observation. The policy network has 228 hidden units and 5 prediction categories, and was trained for 700 iterations of PPO.

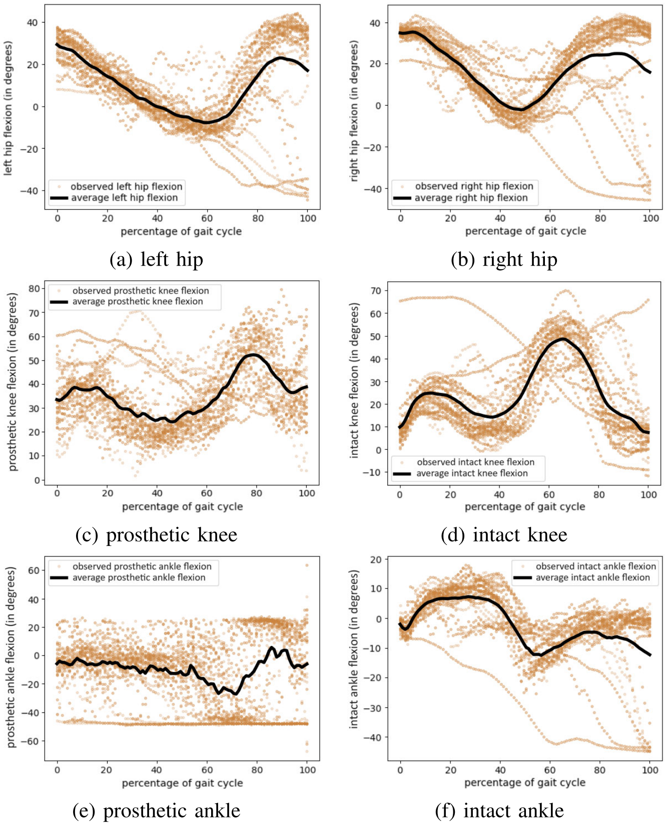
case for the prosthetic ankle, for which there is also a more prominent variation in the emerging patterns.

2) Reduced Observation: Figure 5 shows the instantaneous and average hip, knee, and ankle joints angles for both legs of the agent during 50 gait cycles when a reduced observation state is used. Figures 5a and 5c show that the hip and knee of the prosthetic leg have a clear pattern, but with a much higher variance when compared to the hip and knee of the policy network trained with the complete observation state. Figure 5e shows that there is not a clear trend in the prosthetic ankle of the agent during the gait cycle. This is, however, similar to the results obtained for the policy network trained with the complete observation state (see Figure 4e). Also for the intact joints (hip, knee, ankle) of the agent, there is a high variation during the gait cycle (Figures 5b, 5d, and 5f). This is, however, very different from the clear pattern observed when using the complete observation (see Figure 4, right). The DRL architecture with a reduced observation state does not learn a policy that streamlines the actions performed by the agent to fit a particular gait pattern. Rather, it learns a policy robust enough to perform locomotion but at a cost of human-like movement and decreased similarity to the imitation data.

3) Augmented Observation: Figure 6 shows the instantaneous and average hip, knee, and ankle joints angles for both legs of the agent during 50 gait cycles when an augmented observation state is used. The results show a clear trend in the hips' and knees' angles during the gait cycle. The results also show little variation in the values of the hips' and knees' angles during the gait cycles in comparison to the gait obtained with the reduced observation state. However, there is more variation compared to the policy network trained with the complete muscle information. The intact ankle (Figure 6f) has some variance in the emerging pattern, while the prosthetic ankle does not show a straightforward gait pattern. Similar to the results obtained for the prosthetic ankle trained with the complete observation state (Figure 4e), the emerging pattern contains high variance. There is, however, still a similar trend for both prosthetic ankles between 60 and 80% of the gait cycle.

C. Gait Performances

Table VIII reports the RMSE, symmetry angle, and trend symmetry between the hips, knees, and ankles joints, for the imitation data and for the DRL architecture trained with complete, reduced, and augmented observation state. The most symmetrical DRL architecture for the hip and ankle joints is the one with the complete observation state, i.e., the one with the lower trend symmetry (0.04 and 0.02, respectively). For the knee joint, the most symmetrical DRL architecture is the one with the augmented observation state (0.05). Overall, the architecture with the augmented observation has slightly worse performances than the one with the complete observation. The architecture with the reduced observation state produces the worst symmetry scores.



(a) left hip (b) right hip

(c) prosthetic knee (d) intact knee

(e) prosthetic ankle (f) intact ankle

Fig. 6. Angles of the hips, knees, and ankles during the gait cycle using the augmented state observation. The policy network has 228 hidden units and 5 prediction categories, and was trained for 700 iterations of PPO.

TABLE VIII

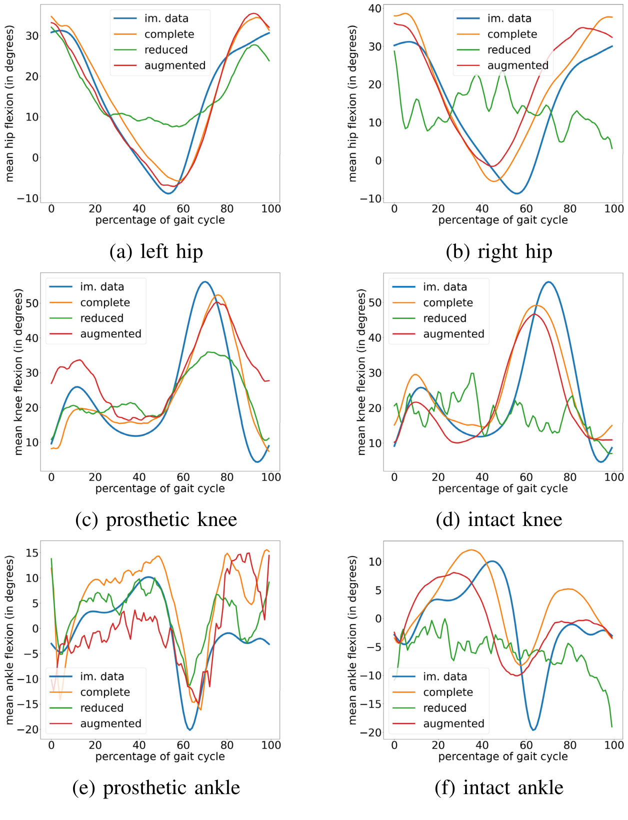
Gait Performances During the Gait Cycle. The Comparison Is Made Between the Joints (Hips, Knees, and Ankles) for the Imitation Data and for the DRL Architecture Trained With the Different Observation State

Joint	Symmetry Metric	DRL Architecture			
		imitation data	complete obser.	reduced obser.	augmented obser.
Hip	RMSE	2.45	6.08	10.56	9.70
	Symmetry Angle	12.64	25.89	18.11	25.46
	Trend Symmetry	0.01	0.04	0.46	0.07
Knee	RMSE	0.46	10.92	9.94	15.11
	Symmetry Angle	0.67	10.03	10.94	16.36
	Trend Symmetry	0.01	0.06	0.45	0.05
Ankle	RMSE	0.31	8.38	10.40	8.30
	Symmetry Angle	2.02	44.13	78.35	69.90
	Trend Symmetry	0.02	0.19	0.37	0.20

TABLE IX

Gait Performances During the Gait Cycle. The Averages Are Compared to the Imitation Data for the Hips, Knees, and Ankle Joints

Joint	Symmetry Metric	DRL Architecture		
		complete observ.	reduced observ.	augmented observ.
Left Hip	RMSE	4.02	8.06	3.87
	Symmetry Angle	19.75	49.31	13.48
	Trend Symmetry	0.02	0.04	0.02
prosthetic Knee	RMSE	7.44	9.45	11.99
	Symmetry Angle	8.99	10.39	12.56
	Trend Symmetry	0.07	0.04	0.10
prosthetic Ankle	RMSE	8.44	5.36	8.02
	Symmetry Angle	57.13	42.80	62.96
	Trend Symmetry	0.16	0.07	0.12
Right Hip	RMSE	6.48	16.70	9.54
	Symmetry Angle	24.55	34.40	28.27
	Trend Symmetry	0.04	0.25	0.08
intact Knee	RMSE	7.46	16.69	9.42
	Symmetry Angle	8.58	13.94	10.06
	Trend Symmetry	0.07	0.41	0.08
intact Ankle	RMSE	6.48	8.61	6.87
	Symmetry Angle	44.61	55.00	36.33
	Trend Symmetry	0.20	0.33	0.17



(a) left hip (b) right hip

(c) prosthetic knee (d) intact knee

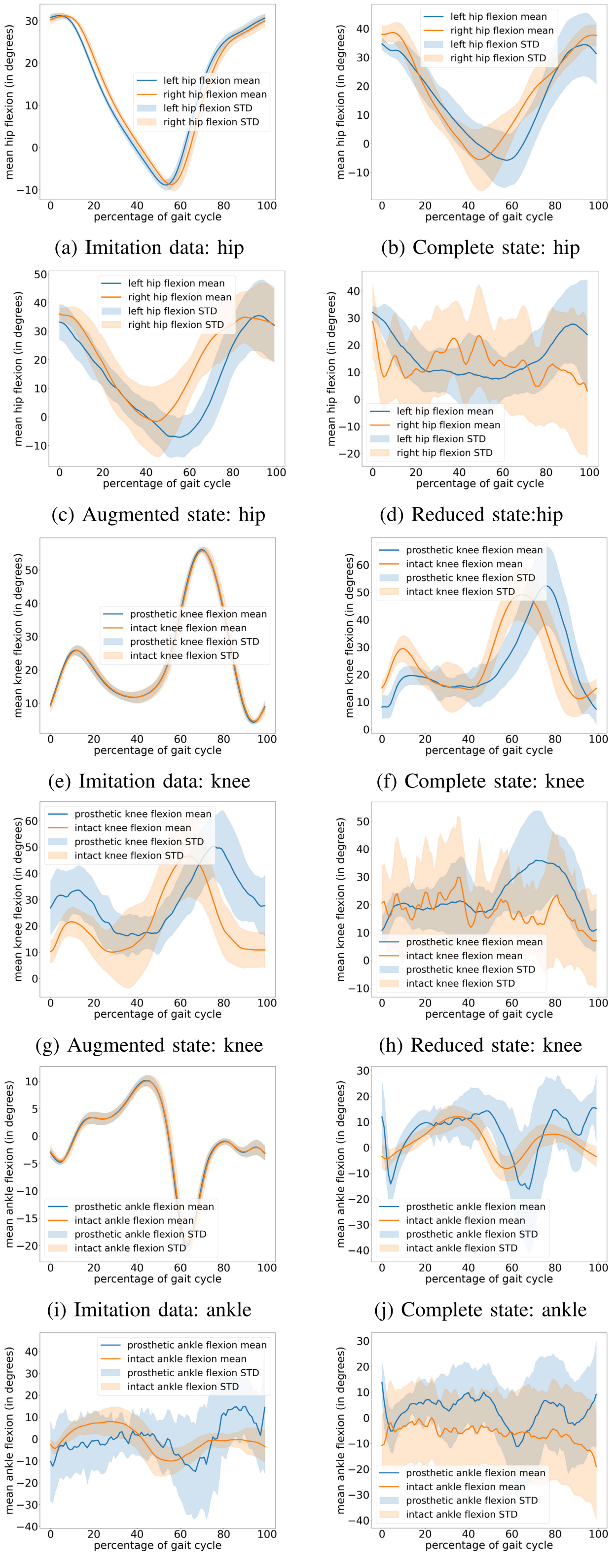
(e) prosthetic ankle (f) intact ankle

Fig. 7. Average angles of the hips, knees, and ankles during the gait cycle using the complete, reduced and augmented observation models, as well as the imitation data.

Table IX reports the average RSME, symmetry angle, and trend symmetry for the hip, knee, and ankle joints for the DRL architecture trained with complete, reduced, and augmented observation state, with respect to the imitation data, as also detailed in Figures 7 and 8. It is possible to note that the DRL algorithm with complete observation state generates a gait with better hip and ankle symmetry than the augmented observation state model. However, the algorithm with the augmented observation state has a better trend symmetry for the ankle joint with respect to the imitation data than when the complete observation state is used. The cases with complete and augmented observation have relatively high trend symmetry with the imitation data for the hip and knee joints. There is, however, an observable decrease in symmetry with the imitation data for the ankles.

D. Actuators' and Muscles' Analysis

Additional analysis was done on the DRL architecture trained with the augmented observation state. This includes



(a) Imitation data: hip (b) Complete state: hip

(c) Augmented state: hip (d) Reduced state:hip

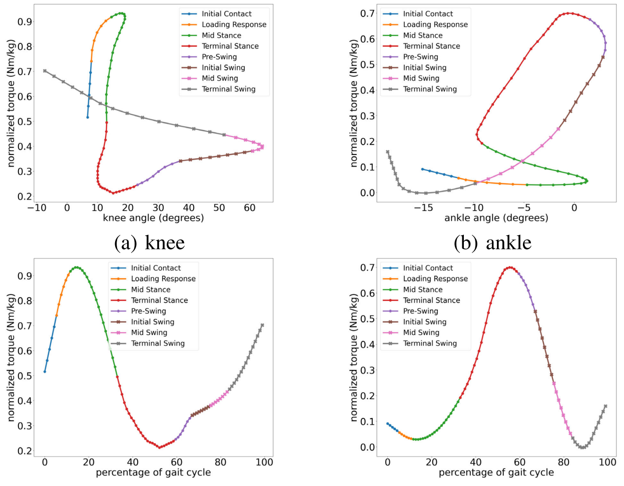
(e) Imitation data: knee (f) Complete state: knee

(g) Augmented state: knee (h) Reduced state: knee

(i) Imitation data: ankle (j) Complete state: ankle

(k) Augmented state: ankle (l) Reduced state: ankle

Fig. 8. Mean and standard deviation during the gait cycles using the different observation states.



(a) knee (b) ankle

(c) knee (d) ankle

Fig. 9. Knee and ankle actuators. (top) Torques (normalized per body weight) with respect to the joint angle during a gait cycles. (bottom) Torques (normalized per body weight) with respect to the percentage of gait cycles.

investigating the required torque/stiffness of the prosthetic knee and ankle joints, as well as observing the forces of the agent's muscles during the gait cycle. For this analysis, the gait cycle is divided into eight phases: initial contact, loading response, mid stance, terminal stance, pre-swing, initial swing, mid swing, and terminal swing. The first five form the stance phase, while the latter three form the swing phase.

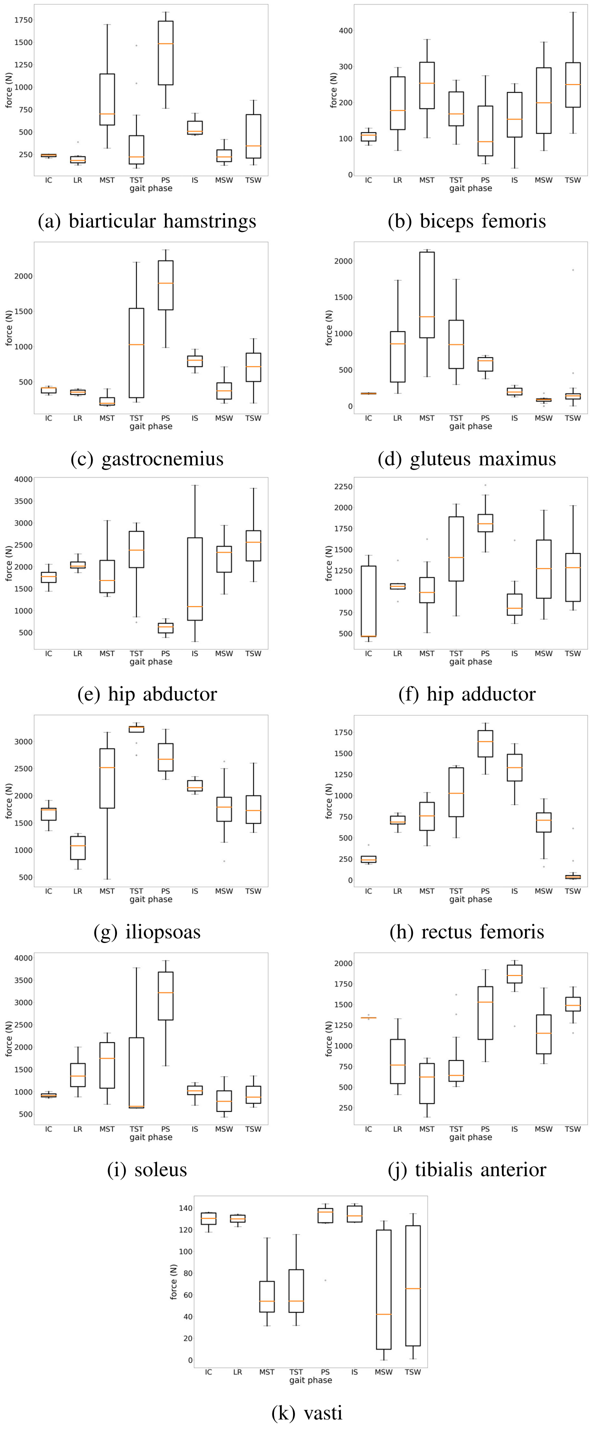
Figure 9 (top) shows the knee/ankle actuator torques (normalized per body weight) with respect to the knee/ankle joint angles during a gait cycle, and Figure 9 (bottom) shows the normalized knee/ankle torques along the gait cycle. The gait phases are also highlighted to give a better description of the torques. These torque values have been filtered by using a Savitzky-Golay filter with a polynomial order of 3 because, due to the lack of constraint on the output of the neural network, there is freedom for the network to result in highly fluctuating activations of the actuators. Moreover, to present a clearer depiction of the knee and ankle torques, a moving average filter was applied to the instantaneous normalized torque waveform using a window size of 0.11 s. Linear extrapolation was used to fill the truncated parts of the waveform that resulted from the moving average filter.

Figure 10 shows the box plots of the muscle forces during the phases of a gait cycle for the 11 muscles in the intact leg of the agent. The combination of the actuators' torques and the muscles' forces during the gait can give insights in the practicality of using the DRL architecture with augmented observation state in real scenarios.

V. Discussion

The results of the DRL architecture trained using the complete, reduced, and augmented state observations were presented in different ways. First, the average reward of 50 episodes of the trained architecture was obtained. To further evaluate the models, the symmetry of both legs, as well as the symmetry with the imitation data was calculated.

While training the PPO model using 2, 3, and 5 prediction categories, training with the complete observation state generated the best performance (obtained reward) for the



(a) biarticular hamstrings (b) biceps femoris

(c) gastrocnemius (d) gluteus maximus

(e) hip abductor (f) hip adductor

(g) iliopsoas (h) rectus femoris

(i) soleus (j) tibialis anterior

(k) vasti

Fig. 10. Muscle force during phases of gait cycle.

agent. The augmented observation state also yielded adequate performance on the locomotion task. It even outperformed the complete observation state model in one instance (2 prediction categories). There was, however, a drop-off in the performance of the model when using the reduced observation for the agent. This large disparity in rewards generated from training with the reduced state and the other observation states suggests that the muscle information or at least an approximation of it is important in the state representation of the agent in order to adequately perform locomotion with the DRL framework.

Over 50 episodes of the simulation, the hip, knee, and ankle joint for the intact leg did not have a clear trend during the gait cycle (Figure 5) unlike in the prosthetic leg when trained with the reduced state. There was a high variation in the flexion progression for different gait cycles. The lack of coordination of the intact leg by the reduced state model is further reflected by the lower symmetry with the imitation data of the hip, knee, and ankle for that leg. For this reduced state model, the symmetry for the prosthetic leg and the imitation data is much higher than the symmetry of the intact leg. This result is expected because there are more muscles in the intact leg and the reduced state observation does not have information on these values in its description. The learning agent thus learns a robust way of maximizing rewards. This learned policy for the reduced state model is not well-correlated with the imitation data and causes the agent to have a less symmetrical and human-like gait in comparison to the model trained using the complete and augmented observation.

The stance phase represents about 60% of the gait cycle [24]. This is true for the imitation data (Figure 7f). It is also observable in the prosthetic leg of the model trained with the complete, augmented, and reduced observation states. There is a steep increase in the hip flexion from approximately 60% of the gait cycle for the rest of the gait cycle (the swing phase). This corresponds to the agent's foot leaving the ground, the hips swinging the leg forward, and finally, the contact of the feet and the ground. In the intact leg, however, the stance phase occurs on average, for only 50% of the gait cycle. This is due to the quicker motion of the prosthesis in comparison to the intact leg. During the mid and terminal swing phases, the prosthetic leg is able to swing faster than the intact leg. This is consistent for both the complete and augmented observation state models and can be due to a number of factors. One reason for this could simply be the lighter mass of the prosthesis in the simulation. The imitation data used to train the model was obtained from non-disabled adults. Transfemoral amputee users of mechanical and microprocessor-controlled prostheses have been observed to have such asymmetry in their gait [3], [25].

Overall, the variance of the hip, knee, and ankle flexion for the 50 episodes of the simulation of the model trained with the augmented state is much more similar to that of the model trained using the complete state. The model has a more natural gait pattern when trained by augmenting the missing muscle information with a pre-trained neural network than when trained without the muscle information; this again, indicates the significance of the muscle data, or an approximation of it, in the observation state to generate a healthy gait pattern.

A. Applicability to Real-World

1) Actuators' Torques/Stiffnesses: Understanding the knee and ankle joints stiffnesses achieved in the simulations could help in further analyzing the gait of the agent, as well as improving the prosthesis mechanical and control design. In this study, the normalized maximum observed normalized knee and ankle torque during the gait of the agent are 0.93 Nm/kg and 0.70 Nm/kg, respectively. The values of these joint torques and the corresponding stiffnesses lie within a reasonable magnitude for the knee and ankle joint of people with lower-limb amputation during a gait [25], [26], [27]. It is, however, unconventional for the observed maximum knee torque to be greater than the ankle torque. This is because the ankle should act as a lever for propulsion and is normally the largest contributor to positive work [28]. Hence, this result also shows a relatively low stiffness in the prosthetic ankle joint. This could lead to more work done by the knee actuator and hip of the prosthesis' user, in order to maintain a normal gait pattern. This can further be improved by methods such as reward shaping to encourage more work being done by the ankle actuator of the prosthesis.

2) Muscle Force: The muscles' forces of the agent during a gait cycle of the simulation are achievable within the physical constrained of the OpenSim simulation environment. However, it should be noted that the `gait1415+2` model is a simplified abstraction of more realistic OpenSim musculoskeletal models with more DOFs and more musculotendon units [2]. Therefore, a comparison with experimental data would not be plausible. Nevertheless, general consideration can be drawn to confirm the validity of the proposed DRL method.

During the initial contact phase of the gait cycle, the tibialis anterior muscle generates a high amount of force (Figure 10j). This is expected because the tibialis anterior muscle's primary function is ankle flexion and extension and in the initial contact phase of the gait cycle, there is weight acceptance by the ankle. Similarly, there is an increase in the muscle force in the pre-swing phase, between the terminal stance and the initial swing phases. This gait phase is a transition between the stance and swing phase of the gait cycle and the foot is pushed and lifted off the ground. There is an increase in force for the muscles responsible for ankle flexion and extension (tibialis anterior and soleus) in this gait phase (Figures 10j and 10i).

At the terminal stance phase, the iliopsoas muscle (Figure 10g) which contributes to the flexion of the hip joint records its maximum force with a median of over 3000 N. Similarly, the rectus femoris muscle (Figure 10h) also contributes to the flexion of the hip and increases its force in the terminal stance and pre-swing gait phases, in preparation for the swing phase of the gait cycle. Conversely, the generated force by the gluteus maximum (Figure 10d) reduces in the terminal stance and pre-swing phases and decreases further during the swing phase of the gait cycle. The gluteus maximum's primary function is hip extension. This is why, to achieve locomotion, this muscle's activation decreases during the swing phase of the gait, in order to achieve the necessary flexion in the hip for the swinging motion of the leg. This is also observed in the biarticular hamstrings (Figure 10a) which contribute to the flexing of the knee joint. For this muscle, much lower force values are observed in the swing phase of the gait.

In the early stages of the swing phase, there is a need for the knee's flexion to increase in order to have sufficient clearance from the ground to swing the leg forward. The biarticular hamstrings (Figure 10a) also contribute to the flexion of the knee and it can be observed that there is a significant increase in the force generated by the muscle during the pre-swing phase. This is however not replicated in the biceps femoris which also contributes to the flexion of the knee. A reason for this is likely that the DRL algorithm learns a policy that mostly utilizes the biarticular hamstrings for the flexion of the knee but not the biceps femoris. Reward shaping can also be used to reduce the over-reliance on specific muscles such as the biarticular hamstrings by incorporating some biomechanical information in the calculation of rewards.

3) Simulations for the Control of Lower-Limb Prostheses: The empirical findings of this study demonstrate that it is possible to train a transfemoral amputee agent without muscle information to accomplish a locomotion task, but with significant deviations from the intended gait pattern, particularly in the intact leg. This is evident in the higher asymmetry of the gait and lower rewards obtained. However, when the agent was trained using a prediction of the muscle data (force, length, and velocity), it achieved significantly higher episodic rewards. These results highlight the importance of including muscle data in learning locomotion for transfemoral amputee models. Therefore, with the final goal of using DRL for the control of lower-limb prostheses, the muscle data, which are complex to measure and process, could be predicted in simulation [18].

4) Implementation: The proposed method of augmenting the muscle information by means of a neural network is not computationally expensive and has the potential of being used in real-time applications.

B. Limitations and Outlook to the Future

1) Reward Shaping: This study highlights the advantages of incorporating supplementary rewards. The study also identifies that the current actuator activation pattern generates undesired bursts of force, which could be improved through reward shaping techniques to reduce variability. Additionally, future research can explore reward shaping based on human knowledge of muscle activations during gait, aiming to achieve even more realistic locomotion for the agent.

2) Incorporate Second-Order Information: This study utilized positions and velocities as observation states to train the muscle predictor network, excluding accelerations. The limitation of only having first-order information may impact the network's ability to predict muscle forces accurately, as they relates to accelerations. Further research can explore the potential benefits of incorporating second-order data for the muscle prediction and its impact on gait generation.

3) Continuous Action Space: The study findings showed a decrease in rewards as the number of prediction categories of the policy network increased. Further research can explore the relationship between this expanded action space and the performance of models trained with reduced/augmented observation states. Optimizing a policy network with a continuous action space can be also investigated.

VI. Conclusion

This study aimed to train a transfemoral amputee model to walk using a state of the art DRL algorithm while observing a reduced number of states. The findings indicate that restricting access to the agent's muscle information significantly hampers its ability to exhibit human-like locomotion. However, by supplementing the reduced state with predicted muscle information from a pre-trained neural network, the reward and gait symmetry of the agent improved. This technique enabled the agent to perform similarly to when it had access to complete observations, including muscle information. These results highlight the importance of muscle information in achieving a natural walking pattern for the trained model.

References

[1] D. L. Robinson et al., "Load response of an osseointegrated implant used in the treatment of unilateral transfemoral amputation: An early implant loosening case study," *Clin. Biomech.*, vol. 73, pp. 201–212, Mar. 2020.

[2] V. Raveendranathan, V. G. M. Kooiman, and R. Carloni, "Musculoskeletal model of osseointegrated transfemoral amputees in OpenSim," *PLoS ONE*, vol. 18, no. 9, Sep. 2023, Art. no. e0288864.

[3] V. J. Harandi et al., "Gait compensatory mechanisms in unilateral transfemoral amputees," *Med. Eng. Phys.*, vol. 77, pp. 95–106, Mar. 2020.

[4] V. J. Harandi et al., "Individual muscle contributions to hip joint-contact forces during walking in unilateral transfemoral amputees with osseointegrated prostheses," *Comput. Methods Biomech. Biomed. Eng.*, vol. 23, no. 14, pp. 1071–1081, Oct. 2020.

[5] L. Frossard, L. Cheze, and R. Dumas, "Dynamic input to determine hip joint moments, power and work on the prosthetic limb of transfemoral amputees: Ground reaction vs knee reaction," *Prosthetics Orthotics Int.*, vol. 35, pp. 141–149, Jun. 2011.

[6] R. Dumas, R. Brånemark, and L. Frossard, "Gait analysis of transfemoral amputees: Errors in inverse dynamics are substantial and depend on prosthetic design," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 679–685, Jun. 2017.

[7] D. T. Davy and M. L. Audu, "A dynamic optimization technique for predicting muscle forces in the Swing phase of gait," *J. Biomech.*, vol. 20, no. 2, pp. 187–201, Jan. 1987.

[8] L. De Vree and R. Carloni, "Deep reinforcement learning for physics-based musculoskeletal simulations of healthy subjects and transfemoral prostheses' users during normal walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 607–618, 2021.

[9] A. J. C. Adriaenssens, V. Raveendranathan, and R. Carloni, "Learning to ascend stairs and ramps: Deep reinforcement learning for a physics-based human musculoskeletal model," *Sensors*, vol. 22, no. 21, p. 8479, Nov. 2022.

[10] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction* (Adaptive Computation and Machine Learning Series), 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[11] A. Ananthakrishnan, V. Kanakiva, D. Ved, and G. Sharma, "Automated gait generation for simulated bodies using deep reinforcement learning," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 90–95.

[12] L. Liu and J. Hodgins, "Learning to schedule control fragments for physics-based characters using deep Q-learning," *ACM Trans. Graph.*, vol. 36, no. 4, p. 1, Jul. 2017.

[13] L. C. Melo and M. R. O. A. Maximo, "Learning humanoid robot running skills through proximal policy optimization," in *Proc. Latin Amer. Robot. Symp.*, 2019, pp. 37–42.

[14] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," 2017, *arXiv:1502.05477*.

[15] S. Lee, M. Park, K. Lee, and J. Lee, "Scalable muscle-actuated human simulation and control," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–13, Aug. 2019.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[17] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–35, Mar. 2018.

[18] M. Sartori, D. G. Llyod, and D. Farina, "Neural data-driven musculoskeletal modeling for personalized neurorehabilitation technologies," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 5, pp. 879–893, May 2016.

[19] R. Carloni, R. Luinge, and V. Raveendranathan, "The gait1415+2 OpenSim musculoskeletal model of transfemoral amputees with a generic bone-anchored prosthesis," *Med. Eng. Phys.*, vol. 123, Jan. 2024, Art. no. 104091.

[20] D. G. Thelen, "Adjustment of muscle mechanics model parameters to simulate dynamic contractions in older adults," *J. Biomech. Eng.*, vol. 125, no. 1, pp. 70–77, Feb. 2003.

[21] M. H. Schwartz, A. Rozumalski, and J. P. Trost, "The effect of walking speed on the gait of typically developing children," *J. Biomech.*, vol. 41, no. 8, pp. 1639–1650, 2008.

[22] R. A. Zifchock, I. Davis, J. Higginson, and T. Royer, "The symmetry angle: A novel, robust method of quantifying asymmetry," *Gait Posture*, vol. 27, no. 4, pp. 622–627, May 2008.

[23] S. J. Crenshaw and J. G. Richards, "A method for analyzing joint symmetry and normalcy, with an application to analyzing gait," *Gait Posture*, vol. 24, no. 4, pp. 515–521, Dec. 2006.

[24] M. A. Laribi and S. Zeghloul, "Human lower limb operation tracking via motion capture systems," in *Design and Operation of Human Locomotion Systems*, M. Ceccarelli and G. Carbone, Eds. New York, NY, USA: Academic, 2020, pp. 83–107.

[25] S. Di Paolo et al., "Longitudinal gait analysis of a transfemoral amputee patient: Single-case report from socket-type to osseointegrated prosthesis," *Sensors*, vol. 23, no. 8, p. 4037, Apr. 2023.

[26] B. M. M. Gaffney et al., "Osseointegrated prostheses improve balance and balance confidence in individuals with unilateral transfemoral limb loss," *Gait Posture*, vol. 100, pp. 132–138, Feb. 2023.

[27] B. Welke, C. Hurschler, M. Schwarze, E. Jakubowitz, H.-H. Aschoff, and M. Örgel, "Comparison of conventional socket attachment and bone-anchored prosthesis for persons living with transfemoral amputation–mobility and quality of life," *Clin. Biomech.*, vol. 105, May 2023, Art. no. 105954.

[28] D. J. Farris and G. S. Sawicki, "The mechanics and energetics of human walking and running: A joint level perspective," *J. Roy. Soc. Interface*, vol. 9, no. 66, pp. 110–118, Jan. 2012.