IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 August 2023, accepted 16 September 2023, date of publication 19 September 2023, date of current version 25 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317183

RESEARCH ARTICLE

End-to-End High-Level Control of Lower-Limb Exoskeleton for Human Performance Augmentation Based on Deep Reinforcement Learning

RANRAN ZHENG[1], ZHIYUAN YU[2], HONGWEI LIU[2], JING CHEN[2], ZHE ZHAO[2], AND LONGFEI JIA[2]

[1]School of Aerospace Engineering, Beijing Institute of Technology, Beijing 100081, China
[2]Beijing Institute of Precision Mechatronics and Controls, Beijing 100076, China

Corresponding author: Ranran Zheng (chephilor@163.com)

ABSTRACT This paper proposes a novel end-to-end controller for the lower-limb exoskeleton for human performance augmentation (LEHPA) systems based on deep reinforcement learning (E2EDRL). The model-free controller contains two control levels: the high-level control responsible for end-to-end human motion intention recognition based on the exoskeleton state signals and the human-exoskeleton interaction (HEI) force signals by deep neural network predictor, and the low-level control for motion tracking by joint PD controllers. The deep neural network predictor does not require complex kinematic calculations that are inevitable in conventional human motion intention recognition methods. We execute the learning process in simulation to learn the E2EDRL strategy efficiently and safely by constructing a novel multibody simulation environment and proposing its specific hybrid inverse-forward dynamics simulation method. The passive mode (all joints remain unpowered) is introduced as a benchmark for comparison purposes. A novel performance assessment method based on HEI forces is put forward to evaluate the E2EDRL strategy quantitatively. The global ratio of the HEI forces in the E2EDRL strategy relative to those in the passive mode is as low as 0.65. The global reduction of the HEI forces demonstrates the superior control performance of the E2EDRL strategy.

INDEX TERMS Lower-limb exoskeleton for human performance augmentation, end-to-end human motion intention recognition, deep reinforcement learning.

I. INTRODUCTION

The lower-limb exoskeleton for human performance augmentation (LEHPA) system is a particular type of wearable robot that is positioned parallel to the pilot/wearer and augments his strength and endurance by transferring the payload weight to the ground [1], [2], [3]. Integrating human intelligence with robot power, the LEHPA system gains a great advantage over other legged robots in adapting to unstructured environments and exhibits a promising prospect in some applications, for example, military, firefighting, disaster relief, construction, manufacturing, etc. [4], [5]. Research on LEHPA systems started from the 1960s and revived in the new century after a long period of silence [6], [7]. Some representative LEHPA systems were developed in the recent two decades, e.g. Berkeley Lower Extremity Exoskeleton (BLEEX) [8], Human Universal Load Carrier (HULC) [9], XOS2 [10], Hybrid Assistive Leg (HAL)-5 Type-B [11], Body Extender (BE) [12], etc.; however, none are sufficiently mature to meet the requirements of practical applications.

In order to guarantee wearing comfort, the LEHPA system is expected to be transparent to the wearer; namely, it is desired to move as consistently as possible with the wearer to reduce the resistance. The study on human-exoskeleton

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

coordination control strategies for LEHPA systems is one of the most important issues and has acquired more attention from researchers. In the past two decades, numerous control strategies have been put forward to ameliorate the performance of LEHPA systems [13], [14]. From the perspective of the control architecture, these control strategies for LEHPA systems are generally hierarchical [15] and consist of three control levels: the high-level control for human motion intention recognition [16], the mid-level control for gait phase detection [17], and the low-level control for motion tracking and stabilization purposes [18]. In terms of high-level control, human motion intention recognition can be divided into three categories: the desired joint torque estimation, the desired joint angular velocity estimation, and the desired joint angle estimation [13]. The signals used to recognize the human motion intention can be classified into three types: the signals only collected from the exoskeleton, the signals collected from the pilot, and the human-exoskeleton interaction (HEI) force signals measured at human-exoskeleton interfaces [19].

Control strategies estimating the human motion intention only based on signals from the exoskeleton facilitate the complexity reduction and the reliability enhancement of LEHPA systems as they require no additional measurement from the pilot or the human-exoskeleton interfaces. The most representative control strategy in this category is sensitivity amplification control (SAC) [20] which was originally put forward to control BLEEX [8] and then used in the control schemes of XOS2 [10], HULC [9], and Hydraulic Lower Extremity Exoskeleton Robot (HLEER) [21]. In the SAC strategy, the sensitivity transfer function maps the equivalent HEI torque to the joint angular velocity of the LEHPA system, describing the effect of the equivalent HEI torque on the LEHPA system. In order to attain a large closed-loop sensitivity transfer function without directly measuring the equivalent HEI torque, the inverse dynamic model of the LEHPA system is introduced as positive feedback. Consequently, any parameter error in the dynamic model is also amplified and transferred to the controller output, doing great harm to the control effect. Unfortunately, the fussy system identification process is indispensable to obtain the dynamic model with sufficient accuracy [22].

The signals collected from the pilot include physiological signals and kinematic signals. To control HAL series prototypes [11], [23], [24], several surface electromyography (sEMG) signals are collected to estimate the musculoskeletal moment of the pilot, which is amplified and then combined with the dynamic model of the swinging shank to generate the knee torque commands. The Hanyang EXoskeletal Assistive Robot (HEXAR) series prototypes utilize muscle stiffness sensors [25] and muscle circumference sensors [26], [27], [28] to estimate targeted joint angles. Physiological signals reflect the human motion intention directly without information loss and delay, but unfortunately, they are easily influenced by noises or signals from muscles adjacent to electrodes, not to mention their complex calibration procedure. In the subsequent hybrid control strategy of BLEEX [29], human joint angles are calculated by using human kinematic signals collected from seven clinometers mounted on the pilot trunk and left and right thighs, shanks, and feet and then used as the targets of the joint-level motion tracking controllers during the stance phase. In the Nanyang Technological University Lower Extremity Exoskeleton (NTU-LEE) system consisting of an inner exoskeleton attached to the human and an outer exoskeleton for load support [30], [31], the human movements are measured by the inner exoskeleton to implement the master-slave control for the outer exoskeleton system. However, these above sensors not only require careful design to enable them to be fastened to the wearer securely but also increase the time taken to don and doff exoskeletons.

Control strategies recognizing the human motion intention based on HEI force signals have been increasingly popular in recent years. The BE exoskeleton acquires the desired joint velocities of the swinging leg by the HEI force signal measured from the six-axis load cell mounted on the snowboard binding [12]. The HEI force signal in the HEXAR-CR50 system is collected by a multi-axis force/torque sensor mounted in the harness module to calculate the equivalent HEI torques by means of the Jacobian matrix. The equivalent HEI torques are then used to generate the joint torque commands [32]. The Harbin Institute of Technology Load-carrying EXoskeleton (HIT-LEX) system collects the HEI forces at the back and swinging foot to calculate the expected velocities of the kinematic terminals of the stance and swinging legs. The expected velocities are combined with the Jacobian matrix to calculate the expected angular velocities of driven joints [33]. The literature [34] proposes to obtain the desired joint positions of the stance leg by minimizing the integral of the HEI force at the back. The Harbin Institute of Technology's exoskeleton (HEXO) system uses HEI forces at the backpack and swinging foot to estimate the desired trajectories of kinematic terminal points, which are then used to obtain the desired joint angles through kinematics calculation for low-level motion tracking of the stance leg and swing leg respectively [35], [36]. In ECUST Lower-extremity Exoskeleton (ELE-ROBOT) [37], the HEI force at the back is measured by two force sensors mounted vertically on the trunk and used to estimate desired joint torques of the stance leg during the single support phase through geometric calculation.

Deep reinforcement learning (DRL) is an interactive machine learning paradigm integrating deep neural networks [38], [39] into the conventional reinforcement learning (RL) framework [40]. With compact representations and robust generalizations, deep neural networks can scale RL up to Markov Decision Process (MDP) problems with high-dimensional and continuous action spaces, offering a new model-free perspective on controller development for complex dynamic systems. DRL has obtained remarkable achievements in the physics-based character animation [41], [42], [43] and the locomotion control of legged robots [44]

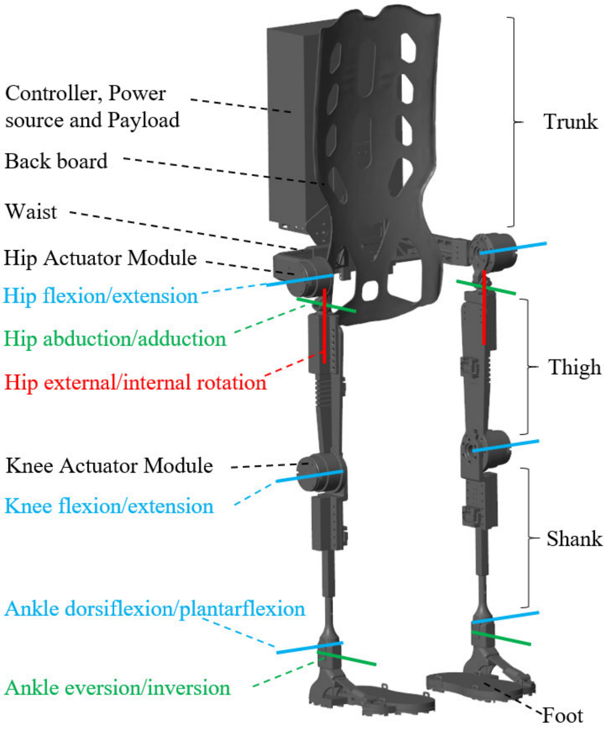


FIGURE 1. Prototype of our LEHPA system.

such as the quadruped [45], [46], [47], [48], [49], the biped [50], [51], [52], [53], and the humanoid [54], [55], [56], [57], [58], etc. In our previous work [59], the DRL framework is introduced to adapt the sensitivity factors of the primary sensitivity amplification controller to ever-changing HEI dynamics. However, no effort has been made to apply DRL to the development of model-free locomotion controllers for LEHPA systems.

This paper investigates the end-to-end control based on deep reinforcement learning (E2EDRL) for our LEHPA system shown in Fig. 1. The model-free controller consists of two control levels: high-level control responsible for end-to-end human motion intention recognition from the exoskeleton signal and the HEI force signal by neural network predictor, and low-level control for motion tracking by joint PD controllers. The main contributions of this paper can be summarized as follows.

1) This paper presents a novel approach to designing a control framework for the LEHPA system without needing any kinematic or dynamic model.
2) This paper proposes a new multibody simulation environment and its corresponding hybrid inverse-forward dynamics simulation method to train the agent.
3) This paper develops a new performance assessment method based on HEI forces to evaluate the control effect of our proposed E2EDRL controller quantitatively.

The remainder of this paper is organized as follows. A detailed description of our control framework is presented in Section II. Section III contains the training in simulation followed by the discussion of results in Section IV. Finally, the conclusion is provided in Section V.

II. END-TO-END HIGH-LEVEL CONTROL BASED ON DEEP REINFORCEMENT LEARNING

An MDP is usually defined by a five-element tuple $(\mathcal{S}, \mathcal{A}, p, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space of the MDP; $\mathcal{A}$, the action space of the MDP; $p(s', r|s, a)$, the environment dynamics specifying a conditional probability distribution for each choice of s and a; $\mathcal{R}$, the reward space, a continuous set of possible rewards; γ, the discount rate. p is not necessary for model-free learning algorithms. The agent interacts with the environment at each of a sequence of discrete time steps. At each time step t, the agent observes the environment state $S_t \in \mathcal{S}$ and on that basis selects an action $A_t \in \mathcal{A}(s)$. One time step later, in part as a consequence of the selected action, the agent observes a new environment state S_{t+1} and receives a numerical reward $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$. The return is defined as the (discounted) sum of the rewards. The agent tries to select the appropriate A_t at each time step t to maximize the expected return.

In the coupled human-exoskeleton system, the human body is powered by the resultant of the human musculoskeletal moment τ_m generated by the pilot muscles and the equivalent HEI torque τ_{HEI} generated by HEI forces at several human-exoskeleton interfaces while the exoskeleton is driven by the resultant of the equivalent HEI torque τ_{HEI} and the actuator torque τ_{act}. The equivalent HEI torque is an assistance to the exoskeleton but a resistance to the wearer. To guarantee wearing comfort, the exoskeleton is desired to move as consistently with the wearer as possible to reduce HEI forces. To this end, the control system is required to accurately recognize and quickly track the human motion intention. By interpreting the human motion intention recognition as an MDP, we propose the E2EDRL strategy whose schematic illustration can be seen in Fig. 2. This control framework has two levels: the high level using deep reinforcement learning to synthesize a policy to predict the human motion intention from the exoskeleton signal and the HEI force signal, and the low level using PD controllers to track the human motion intention. The detailed description of the E2EDRL strategy is as follows.

A. STATE SPACE AND ACTION SPACE

During normal walking, the movements in the frontal plane and transversal plane are rather small and have little dynamic effect on the LEHPA system compared to the movements in the sagittal plane. For the sake of brevity, we neglect the movements in the frontal plane and transversal plane and only consider those in the sagittal plane. It has been a conventional simplification in the research of LEHPA systems.

The state for the MDP problem to be solved is represented as the concatenation of the exoskeleton signal vector X_E and the HEI force signal vector X_{HEI} in the sagittal plane, i.e. $S_t = (X_E, X_{HEI})$. The exoskeleton signal vector X_E consists of 14 components: the trunk pitch angle and angular velocity,

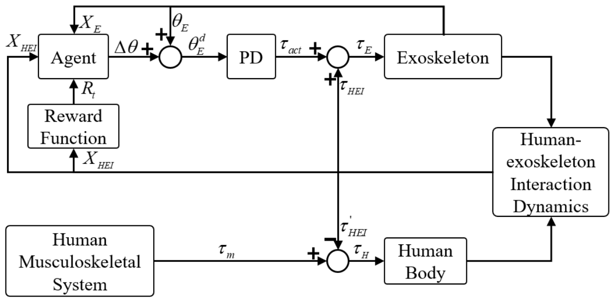


FIGURE 2. Diagram for E2EDRL strategy.

and the joint angles and angular velocities of the left and right hips, knees, and ankles. The HEI force signal vector X_{HEI} includes the HEI forces at the back and left and right thighs and shanks. The HEI force at the back is composed of three components: the pitch torque T_{pitch}, the force along the sagittal axis F_S, and the force along the vertical axis F_V. The HEI force at each lower-limb segment comprises two components, the one along the central axis of the lower-limb segment F_t and the other normal to the central axis F_n. The eight HEI force components at the four lower-limb segments are named F_{nRT}, F_{tRT}, F_{nRS}, F_{tRS}, F_{nLT}, F_{tLT}, F_{nLS}, and F_{tLS} respectively. Fig. 3 depicts all the 11 HEI force components in the signal vector X_{HEI}. The combined representation of the 14 exoskeleton signals and the 11 HEI force signals yields the 25D state space.

The action is represented as the estimated human motion intention, which may take three forms: the target joint angles, the target joint angular velocities, or the target joint torques. In the literature [60], the three different action parameterizations are compared in terms of learning speed, policy robustness, motion quality, and policy query rates. The result shows that choosing the target joint angles for the active joints as the action can greatly improve learning efficiency and control performance for locomotion control problems. Therefore, the target joint angles for the active joints are chosen as the action in this work, leading to a 4D action space. Actually, the pilot leads the exoskeleton to move together by means of HEI forces during locomotion. Even though a human joint angle and corresponding exoskeleton joint angle vary in a large range, the deviation between them remains in a small range. That is to say, the current exoskeleton joint angles provide a hint as to what the target joint angles might be. To improve learning efficiency, our policy learns how to augment the current exoskeleton joint angles instead of directly outputting the target joint angles. The action is represented as the angle augmentations for the active joints, i.e. $A_t = \Delta\theta$. Thus, the target joint angles are the sum of the action and the current exoskeleton joint angles:

$$\theta_E^d = \Delta\theta + \theta_E \tag{1}$$

where θ_E and θ_E^d are the current exoskeleton joint angles and the target joint angles respectively.

To prevent target joint angles from changing dramatically, we limit the maximum augmentation amplitude of all powered joint angles at each time step to $\pi/10$. The target joint angles are passed through a low-pass filter to mitigate undesirable high-frequency movements before being applied to low-level joint PD controllers to generate the following actuator torques for motion tracking:

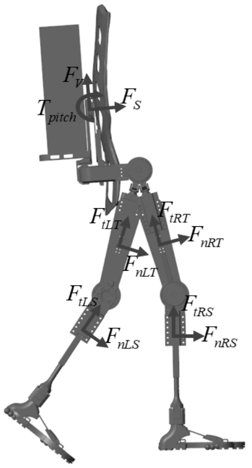


FIGURE 3. HEI forces used in state space.

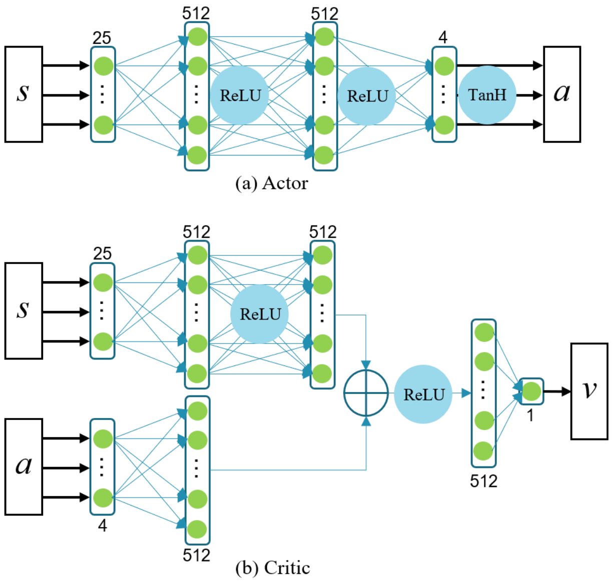
$$\tau_{act} = P(\theta_E^d - \theta_E) - D\dot{\theta}_E \tag{2}$$

where P and D are the gains of low-level joint controllers and $\dot{\theta}_E$ denotes the current joint angular velocities. Note that the high-level policy query rate is far slower than the running rate of low-level joint controllers.

B. LEARNING ALGORITHM AND NEURAL NETWORKS

In this work, Twin Delayed Deep Deterministic Policy Gradient (TD3) [61] which is a model-free off-policy Actor-Critic method is selected as the learning algorithm to learn the E2EDRL strategy. The Actor-Critic architecture combines the advantage of policy gradient methods and that of value function methods. TD3 is modified from Deep Deterministic Policy Gradient (DDPG) [62] that integrates Deep Q-Network (DQN) [63] into Deterministic Policy Gradient (DPG) [64]. TD3 ameliorates DDPG in three aspects: reducing variance by clipped Double Q-Learning that prevents the error from accruing; addressing the coupling of the value and the policy by delaying policy updates until the value estimate has converged; further reducing variance by target policy smoothing regularization strategy in which a SARSA-style update bootstraps similar action estimates. The above three improvements make TD3 more data-efficient than DDPG.

The TD3 agent contains one Actor network and two twin delayed Critic networks that share the same architecture but have separate learnable weight and bias parameters. The architectures of the Actor and Critic networks are illustrated in the two subfigures of Fig. 4 respectively. The Actor network receives the exoskeleton signal and the HEI force signal and outputs the human motion intention. The numbers of neural units in its input layer and output layer are determined by the dimensions of the state space and the



(a) Actor

(b) Critic

FIGURE 4. Architectures of the Actor and Critic networks.

TABLE 1. Notations of F_{ij}.

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$j=1$	T_{pitch}	F_{nRT}	F_{nRS}	F_{nLT}	F_{nLS}
$j=2$	F_S	F_{tRT}	F_{tRS}	F_{tLT}	F_{tLS}
$j=3$	F_V				

action space respectively. Following the input layer are two fully connected 512-neural-unit hidden layers which are both activated by the ReLU (Rectified Linear Unit) function. The output layer is activated by the TanH (hyperbolic tangent) function to limit the range of the final output. The two twin Critic networks receive the state vector and action vector as input and output the value estimates of each state-action pair. The action vector and the state vector are passed through one hidden layer and two hidden layers with 512 neural units respectively, with the first hidden layer of the state activated by the ReLU function. The two 512D vectors originating from the state and action are added up and then activated by the ReLU function. Lastly, the activated vector is fully connected to the one-neural-unit output layer representing the value function.

C. REWARD FUNCTION

It is one of the most distinctive features of reinforcement learning to formalize the goal of the agent by reward signal. The reward function determines the optimization direction of learnable weight and bias parameters of Actor and Critic networks during the training process. It is critical to design a suitable reward function for the MDP because any misspecification of the reward function can have unintended and even dangerous consequences. On one hand, we aim to minimize HEI forces to reduce the obstruction of the exoskeleton to the pilot in the task of LEHPA control; On the other hand, the goal of the agent is to maximize the expected (discounted) return, namely the expected (discounted) sum of the rewards. Thus, the reward function should decrease monotonically with respect to HEI forces. The reward function is designed as the weighted sum of five local reward terms relevant to HEI forces at the five human-exoskeleton interfaces shown in Fig. 3:

$$r = \sum_{i=1}^{5} w_i r_i \tag{3}$$

where r_i denotes the local reward term derived from the HEI force at the i-th human-exoskeleton interface whose contribution to the global reward is determined by its corresponding local weight w_i. All local reward terms take the following form:

$$r_i = \exp(-\sum_j k_{ij}|\varepsilon_{ij}|^2),\ \varepsilon_{ij} = \frac{F_{ij}}{\Delta_{ij}} \tag{4}$$

where ε_{ij} denotes the normalized force component. F_{ij} is the j-th HEI force component at the i-th human-exoskeleton interface, while Δ_{ij} represents the normalization term carefully determined to normalize F_{ij}. The exponent weight k_{ij} determines the contribution of ε_{ij} to the exponent. Table 1 lists the notations of all these above HEI force components.

III. TRAINING IN SIMULATION

A. MULTIBODY SIMULATION ENVIRONMENT

In order to learn the E2EDRL strategy efficiently and safely, we execute the learning process of the TD3 agent in a multibody simulation environment constructed based on the MATLAB/Simscape physical modeling toolbox. As illustrated in Fig. 5, the multibody simulation environment comprises the exoskeleton model, the human body model, HEI models at all the human-exoskeleton interfaces, and the terrain.

1) THE HUMAN BODY MODEL

The human body model is a simplification of the wearer. Given that this work only focuses on the lower-limb movements, the upper limbs are left out. Likewise, the degrees of freedom (DOFs) of lower-limb joints in the frontal and transverse planes are omitted in this work since only the movements in the sagittal plane are considered. Hence, each leg only preserves three DOFs, i.e. the ankle dorsiflexion/plantarflexion, the knee flexion/extension, and the hip flexion/extension. Note that the rigid human body model is more like the human skeleton system from the view of human-exoskeleton interactions. The flexibility of human muscles adjacent to human-exoskeleton interfaces is combined with the harness and is integrated into HEI models. The dimensional and inertial parameters of the human body model are referenced from China national standards "Human dimensions of Chinese adults GB 10000-1988" and "Inertial parameters of adult human body GB/T 17245-2004" respectively.

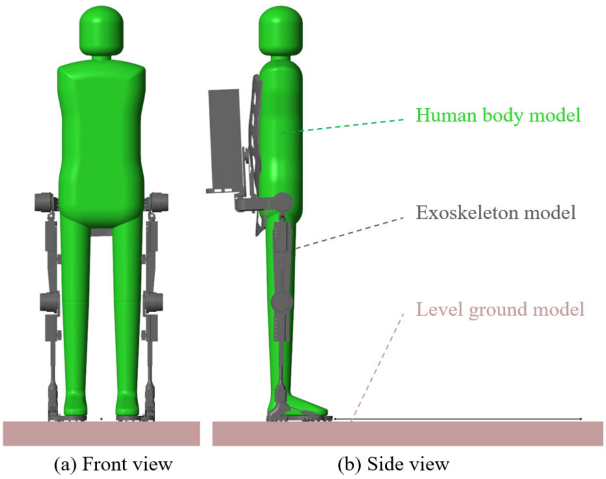


FIGURE 5. The multibody simulation environment.

2) THE EXOSKELETON MODEL

The exoskeleton model is a simplification of the LEHPA system and consists of seven parts, i.e. the trunk and the left and right thighs, shanks, and feet. The trunk is composed of the backboard, the waist, the payload, the control system, and the power source unit. The waist width, thigh length, and shank length are designed to be adjustable to match different wearers. Since this work only focuses on the lower-limb movements in the sagittal plane, we neglect the DOFs in the frontal and transverse planes and only retain DOFs in the sagittal plane. Thus, there are only three DOFs on each leg, namely the ankle dorsiflexion/plantarflexion, the knee flexion/extension, and the hip flexion/extension. The dimensional and inertial parameters of the exoskeleton model are calculated by CAD software.

3) THE HEI MODELS

In this multibody simulation environment, the human body model interacts with the exoskeleton model by means of the HEI models at several human-exoskeleton interfaces, including the back and the left and right thighs, shanks, and feet. In this work, the interactions in the sagittal plane are modeled whereas those in the frontal and transverse planes are ignored.

The HEI at the back human-exoskeleton interface is modeled as the superposition of a torsional spring and a spring-damper system. The torsional spring determines the torque component resulting from the orientation discrepancy at the back human-exoskeleton interface between the human body model and the exoskeleton model while the spring-damper system determines the force component due to the position discrepancy between the two. The HEI at each thigh or shank is modeled as a spring-damper system. The HEI at each foot is modeled as two spring-damper systems placed at the heel and toe respectively. The torsional spring stiffness and damping coefficients are set to 20 Nm/rad and 0.5 Nm·s/rad respectively. Table 2 lists the stiffness and damping coefficients of each spring-damper. These stiffness and damping coefficients are all chosen by experience. The system will oscillate if their values are set too large. However, the exoskeleton model will fall down if their values are set too little. To determine these coefficients, we introduce the passive mode (all active joints remain unpowered), in which the exoskeleton model is only driven by HEI forces. We choose a set of initial values (which are little enough to make the exoskeleton model fall down) for these parameters and then gradually increase them until the exoskeleton model can be driven to move forward together with the human body model.

TABLE 2. Stiffness and damping of spring-damper systems.

Interface	Back	Thigh	Shank	Heel	Toe
Stiffness (N/m)	4e3	1e4	1e4	5e5	5e5
Damping (N·s/m)	4e2	5e2	5e2	5e2	5e2

4) TERRAIN

The terrain in the multibody simulation environment is fixed to the world frame and interacts with two exoskeleton feet at their underneath to generate the ground reaction force to support the weight of the coupled human-exoskeleton system. There are some structured terrains usually employed in the exoskeleton research, e.g. the level ground, stairs of different heights and widths, and slopes of different degrees. In our level walking simulation, the level ground is simplified as a flat plate.

The ground reaction force is the resultant force of the contact forces generated by two rows of Spatial Contact Force blocks respectively placed at the left and right edges of the underneath of each exoskeleton foot. Each contact force generated by a Spatial Contact Force block can be decomposed into two components. The normal component perpendicular to the contact surface is determined by the classical spring-damper model, while the frictional component tangent to the contact surface is determined by the Smooth Stick-Slip law. The stiffness and damping coefficients of each spring-damper model are set to 2e4 N/m and 4e2 N·s/m respectively in the same way as those of HEI models to compute the normal component. Due to the too little initial values, the ground cannot provide sufficient reaction force to support the coupled human-exoskeleton system and prevent it from falling down. We gradually increase them until the generated ground reaction force is sufficient to support the coupled human-exoskeleton system. As for the frictional component, the static and dynamic friction coefficients are set to 0.9 and 0.7 respectively by experience.

B. HYBRID INVERSE-FORWARD DYNAMICS SIMULATION

In this work, a novel hybrid inverse-forward dynamics simulation method is put forward specific to the multibody simulation environment. As can be seen in Fig. 6, the joints of the human body model and the exoskeleton model are modeled as inverse dynamics joints and forward dynamics joints respectively. The inverse dynamics joints receive the joint angle trajectories as inputs whereas the forward dynamics joints receive the joint torques as inputs. During simulation, the human body model is driven by the reference motion and then leads the exoskeleton model to move together by means of HEI forces generated by HEI models, while the controller tries to generate appropriate joint torques to drive the exoskeleton model and reduce HEI forces. The original reference motion is a complete gait cycle of angle trajectories of the trunk pitch and the left and right hips, knees, and ankles collected by the motion capture system (Mtw Awinda, Xsens) from the human walking on the treadmill at 2.8 km/h. The gait cycle lasts about 1.392 s at 240 Hz, including 334 samples in total. It is extended periodically during simulation. In order to extensively explore the state space, reference motions of different walking speeds are acquired by stretching or compressing the gait cycle of the original reference motion. For a certain episode, the specific walking speed is limited to the interval of [2.8 km/h, 5.6 km/h] by randomly selecting the gait cycle in the interval [0.696 s, 1.392 s]. It is worth noting that these resulting reference motions are still physically feasible because the stance foot of the exoskeleton model won't slide along the ground during walking. In contrast to previous methods that directly input the reference motion into the exoskeleton model to calculate the desired joint torques, the hybrid inverse-forward dynamics simulation method demonstrates the dynamic HEI process of the coupled human-exoskeleton system during walking, unveiling the nature of the coordinated human-exoskeleton movement.

C. TRAINING SETUP

The training runs episodically. A sample point is chosen randomly from angle trajectories to concurrently initialize the pitch and joint angles of both the human model and the exoskeleton model at the beginning of each episode. A rollout is then simulated by taking the action selected by the policy at every time step. The time horizon, i.e. the possible longest simulation time of an episode, is set to 5s. In case of excessive explorations of poor states, the early termination mechanism is proposed to cease the current episode and set the remaining rewards to 0. The early termination mechanism is triggered whether the absolute exoskeleton trunk pitch angle value is over $\pi/9$ rad or the vertical discrepancy between the two sides of the back human-exoskeleton interface is more than 0.3 m. Hence, an episode terminates when an early termination occurs or until the simulation time reaches the predetermined time horizon. The simulation rate in the training process is set to 2 kHz. The policy query rate is set to 25 Hz to update the target joint angles every 40 ms, while the low-level joint PD controllers run at the same rate as the simulation.

Table 3 shows the values of local weights, normalization terms, and exponent weights set by experience. Note that it is inappropriate if the normalization term Δ_{ij} is set too large or too small. From Eq. (4) we can obtain the derivative of the local reward r_i with respect to the normalized force component ε_{ij}

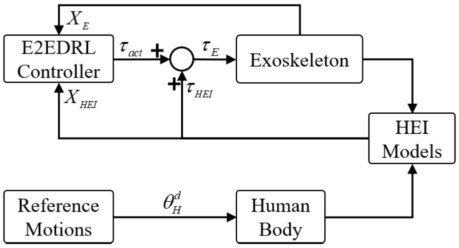


FIGURE 6. Diagram for hybrid inverse-forward simulation.

$$\frac{dr_i}{d\varepsilon_{ij}} = -2k_{ij}\varepsilon_{ij}r_i. \tag{5}$$

Figure 7 illustrates the variations of r_i and $\frac{dr_i}{d\varepsilon_{ij}}$ with ε_{ij}. k_{ij} is set as 1 in the two curves. Note that k_{ij} only changes the values of r_i and $\frac{dr_i}{d\varepsilon_{ij}}$, but does not change their trends. To ensure the learning speed, it is desired to distinguish "good" actions leading to little HEI forces from "bad" actions leading to large HEI forces. That is to say, the normalization term Δ_{ij} should be carefully chosen to make the normalized force component ε_{ij} as close to the range with large $\frac{dr_i}{d\varepsilon_{ij}}$ as possible.

It is obvious from Fig. 7 that $\frac{dr_i}{d\varepsilon_{ij}}$ is close to 0 when ε_{ij} is too little or too great. The extreme point of $\frac{dr_i}{d\varepsilon_{ij}}$ is close to 0.7. Thus, Δ_{ij} should share the same order of magnitude as the maximum absolute value of F_{ij}. To determine the range of each HEI force component, we introduce the passive mode (all joints remain unpowered) as the benchmark. We set Δ_{ij} to about 0.9 times the maximum value of F_{ij} during a whole gait cycle in the passive mode.

The simulation is executed in parallel, with 20 workers running simultaneously on a 20-core Intel Xeon CPU. The Actor and Critic networks are trained on an NVIDIA GeForce RTX 2080 Ti GPU. It takes about 2.5h to finish 800 episodes of simulation.

IV. RESULTS AND DISCUSSION

In this work, the HEI forces at the back, thighs, and shanks are used to evaluate the performance of our E2EDRL controller. Root-mean-square (RMS) values of these HEI forces are chosen as the performance indicator:

$$\bar{F} = \sqrt{\frac{1}{T}\int_0^T F^2 dt} \tag{6}$$

where F denotes a certain HEI force component. T is the time duration. Given that HEI forces are not strictly periodic, we set T to 5 gait cycles.

The passive mode is used as a benchmark for comparison purposes. The normalized ratio of each HEI force component F is defined as the ratio of its RMS value in the proposed E2EDRL strategy to that in the passive mode to evaluate its improvement:

$$\lambda(F) = \frac{\bar{F}_{E2EDRL}}{\bar{F}_{PSV}} \tag{7}$$

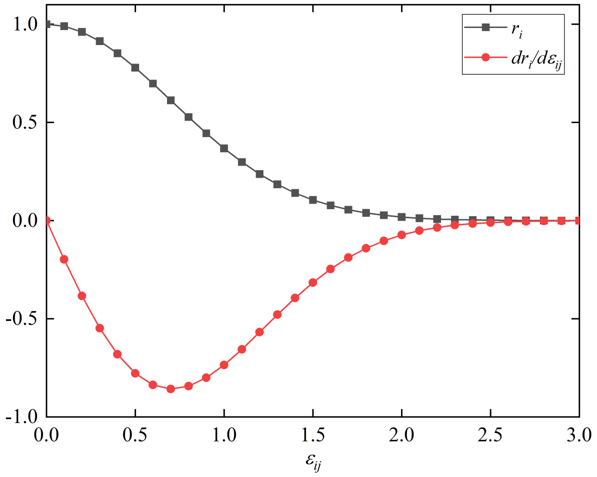


FIGURE 7. Diagram of reward function characteristics.

TABLE 3. Values of reward function parameters.

i	w_i	j	Δ_{ij}	k_{ij}
		1	4.5	0.5
1	0.4	2	250	0.25
		3	300	0.25
2	0.15	1	150	0.75
		2	50	0.25
3	0.15	1	60	0.75
		2	6	0.25
4	0.15	1	150	0.75
		2	50	0.25
5	0.15	1	60	0.75
		2	6	0.25

where $\bar{F}_{E2EDRL}$ and $\bar{F}_{PSV}$ denote the RMS values of F in E2EDRL and the passive mode respectively.

To investigate the effect of the walking speed on the performance improvement, we calculate the normalized ratios of F at five different reference walking speeds, 2.8 km/h, 3.5 km/h, 4.2 km/h, 4.9 km/h, and 5.6 km/h and the weighted average ratio for each HEI force component by:

$$\bar{\lambda}(F) = \sum_{i=1}^{5} \mu_i \lambda_i(F) \tag{8}$$

where $\lambda_1(F)$, $\lambda_2(F)$, $\lambda_3(F)$, $\lambda_4(F)$, and $\lambda_5(F)$ represent the normalized ratios of F at 2.8 km/h, 3.5 km/h, 4.2 km/h, 4.9 km/h, and 5.6 km/h respectively, while μ_i is the weight of $\lambda_i(F)$.

Generally, humans select different walking speeds at different frequencies, selecting walking speeds closer to their self-selected walking speeds more frequently. Thus, the natural choice is to assign greater weight to the walking speed closer to the self-selected walking speed. We chose a stair-like weight set 0.1, 0.15, 0.2, 0.25, 0.3 and allocated an element to the normalized ratio at each specified walking speed according to its speed difference to the self-selected walking speed. We tested the self-selected walking speeds for some subjects between the heights 165 cm and 175 cm on the treadmill and found that their self-selected walking speeds are close to 4 km/h. So, we set the nominal self-selected walking speed to 4 km/h. Consequently, the five specified walking speeds are allocated the weights 0.15, 0.25, 0.3, 0.2, and 0.1 respectively, i.e. $\mu = [0.15, 0.25, 0.3, 0.2, 0.1]$.

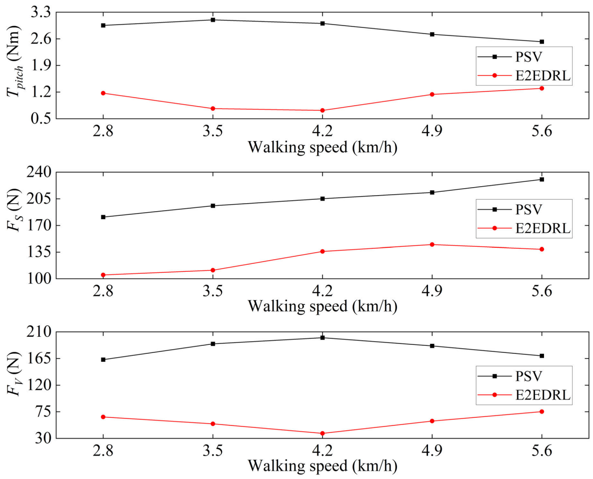


FIGURE 8. HEI force RMS at the back.

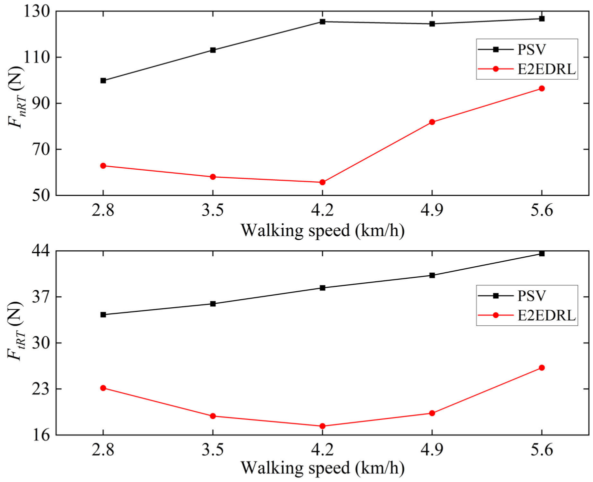


FIGURE 9. HEI force RMS at the right thigh.

To evaluate the comprehensive improvement, the global ratio is defined as the weighted sum of weighted average ratios of HEI force components:

$$\lambda^{\star} = \sum_{i} w_i \sum_{j} k_{ij} \bar{\lambda}(F_{ij}) \tag{9}$$

where w_i and k_{ij} represent respectively the local and exponent weights, which have been used in the reward function expression.

Given that the movements of the two legs of the coupled human-exoskeleton system are symmetrical, it is a rational assumption that each HEI force component at the left leg has the same weighted average ratio as its counterpart at the right

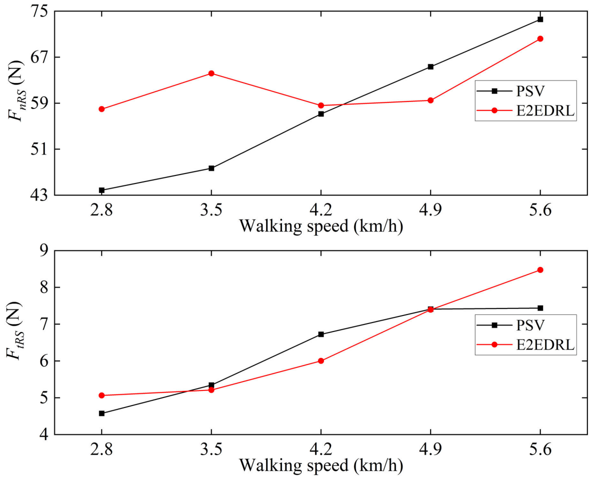


FIGURE 10. HEI force RMS at the right shank.

TABLE 4. Ratios of HEI forces.

F	$\lambda_1(F)$	$\lambda_2(F)$	$\lambda_3(F)$	$\lambda_4(F)$	$\lambda_5(F)$	$\bar{\lambda}_5(F)$
T_{pitch}	0.40	0.28	0.24	0.42	0.51	0.33
F_S	0.58	0.57	0.66	0.68	0.60	0.62
F_V	0.41	0.29	0.19	0.32	0.44	0.30
F_{nRT}	0.63	0.51	0.44	0.66	0.76	0.56
F_{tRT}	0.67	0.53	0.45	0.48	0.60	0.52
F_{nRS}	1.32	1.35	1.03	0.91	0.95	1.12
F_{tRS}	1.11	0.98	0.89	1.00	1.14	0.99

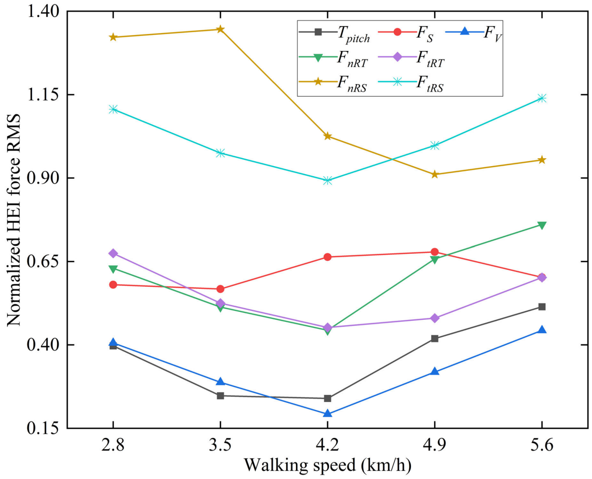


FIGURE 11. Normalized ratios of HEI forces.

leg at the timescale of gait cycles:

$$\begin{cases} \bar{\lambda}(F_{nLT}) = \bar{\lambda}(F_{nRT}) \\ \bar{\lambda}(F_{tLT}) = \bar{\lambda}(F_{tRT}) \\ \bar{\lambda}(F_{nLS}) = \bar{\lambda}(F_{nRS}) \\ \bar{\lambda}(F_{tLS}) = \bar{\lambda}(F_{tRS}) \end{cases} \tag{10}$$

Hence, we only calculate the RMS values and normalized ratios of the HEI force components at the back and right leg for simplicity, leaving out the redundant calculation for the HEI forces at the left leg. The RMS values of HEI force components at the back and right leg at the five selected walking speeds are shown in Fig. 8, Fig. 9, and Fig. 10. It is worth noting that even though the RMS value of each HEI force component in simulation may be different from that in reality due to the reality gap, especially the differences between the interaction models in the multibody simulation environment and the straps in the real world, it does not impact on the comparison.

It can be seen from Fig. 8 and Fig. 9 that the HEI forces at the back and right thigh in E2EDRL are much less than their counterparts in the passive mode. The RMS values of T_{pitch} and F_V in E2EDRL are less than their counterparts in the passive mode respectively, proving that more payload weight is transferred to the ground successfully. The RMS values of F_S, F_{nRT}, and F_{tRT} in E2EDRL are less than their counterparts in the passive mode respectively, meaning that E2EDRL can reduce the misalignment between the pilot and the exoskeleton and improve the motion tracking performance. However, the HEI force at the right shank presented in Fig. 10 shows some differences from the former two HEI forces. At the walking speeds of 2.8 km/h and 3.5 km/h, the F_{nRS} RMS values in E2EDRL are much greater than their counterparts in the passive mode respectively, whereas the F_{nRS} RMS values in E2EDRL are much less than their counterparts in the passive mode at 4.9 km/h and 5.6 km/h. As for F_{tRS}, its RMS values in E2EDRL are greater than those in the passive mode respectively at 2.8 km/h and 5.6 km/h, whereas the value in E2EDRL is less than that in the passive mode at 4.2 km/h. Even though the phenomenon seems obscure, it can still be analyzed from the perspective of dynamics. During the stance phase, the fixed foot acts as the base, and the shank motion is determined by the equivalent HEI torque acting on the passive ankle joint, which is mainly produced by the HEI force at the shank, especially the component normal to the shank link, F_{nRS}. This means that the more payload weight is transferred to the ground, the greater the HEI force at the shank tends to be. Therefore, the HEI force at the shank in E2EDRL should be greater than its counterpart in the passive mode. During the swing phase, the shank motion is determined by the torque acting on the knee joint, which is mainly produced by the knee actuator rather than the HEI force at the shank. Thus, the HEI force at the shank in E2EDRL should be less than its counterpart in the passive mode. Totally, the result in E2EDRL synthesizes the effects of the increase during the stance phase and the decrease during the swing phase. As for F_{nRS}, the increase during the stance phase dominates at low speeds, but is eclipsed by the decrease during the swing phase at high speeds. Regarding F_{tRS}, the increase during the stance phase dominates at speeds close to the nominal self-selected walking speed, but is overwhelmed by the decrease during the swing phase when the walking speed is far away from the nominal self-selected walking speed, whether too slow or too fast.

The normalized ratios of the seven HEI force components at the five selected walking speeds and their corresponding weighted average ratios are listed in Table 4. These normalized ratios are also presented in Fig. 11 for further analysis. They can be divided into three groups according to their value ranges at the five walking speeds: the first group consists of the normalized ratios of T_{pitch} and F_V, which range approximately from 0.2 to 0.5; the second group is made up of the normalized ratios of F_S, F_{nRT}, and F_{tRT} varying around from 0.45 to 0.75; the third group includes the normalized ratios of F_{nRS} and F_{tRS} ranging about from 0.9 to 1.35. Obviously, the further a human-exoskeleton interface is away from unpowered ankle joints, the less the weighted average ratio of each HEI force components at this interface is, except for F_S. Distinctively, $\bar{\lambda}_{F_S}$ is much greater than $\bar{\lambda}_{F_V}$ and $\bar{\lambda}_{T_{pitch}}$, indicating that the existence of the walking speed makes the deviation along the sagittal axis between the pilot and the exoskeleton at the back more difficult to reduce.

Finally, we can acquire the global ratio for the proposed E2EDRL control strategy according to (9), $\lambda^{\star} = 0.65$.

V. CONCLUSION AND FUTURE WORK

This work investigates a deep reinforcement learning framework to learn a novel model-free walking controller for our LEHPA system. The controller estimates human motion intention directly by a deep neural network and needs no kinematic or dynamic model of the LEHPA system. To learn the TD3 agent efficiently and safely, we execute the learning process in simulation by creating a new multibody simulation environment and proposing its corresponding hybrid inverse-forward dynamics simulation method. To evaluate the control effect of the proposed E2EDRL strategy, the passive mode is introduced as a benchmark. The proposed E2EDRL strategy is compared with the passive mode in terms of the HEI forces at the back, thighs, and shanks. The weighted average ratio and global ratio are defined to evaluate each local HEI force component and global HEI forces respectively. The global ratio is 0.65, proving that the proposed E2EDRL strategy effectively reduces the HEI forces and has superior control effect. This research demonstrates the feasibility to design model-free walking controllers for LEHPA systems using deep reinforcement learning.

Several aspects will be involved in future works. First, the E2DRL controller will be further trained on some more terrains to adapt to complex environments. Some typical terrains, for instance stairs of different heights and widths and slopes of different degrees, will be constructed in the multibody simulation environment. Correspondingly, the reference motions of human walking on these terrains will be collected to drive the human body model. Additionally, in order to transfer the learned control strategy from simulation to reality successfully, some measures will be taken to close the reality gap. Finally, the deep reinforcement learning framework will be implemented on our real LEHPA platform to fine-tune the controller in the real-world environment.

REFERENCES

[1] S. Qiu, Z. Pei, C. Wang, and Z. Tang, "Systematic review on wearable lower extremity robotic exoskeletons for assisted locomotion," *J. Bionic Eng.*, vol. 20, no. 2, pp. 436–469, Oct. 2022. [Online]. Available: https://link.springer.com/10.1007/s42235-022-00289-8

[2] S. Viteckova, P. Kutilek, G. de Boisboissel, R. Krupicka, A. Galajdova, J. Kauler, L. Lhotska, and Z. Szabo, "Empowering lower limbs exoskeletons: State-of-the-art," *Robotica*, vol. 36, no. 11, pp. 1743–1756, Nov. 2018. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0263574718000693/type/journal_article

[3] S. Yeem, J. Heo, H. Kim, and Y. Kwon, "Technical analysis of exoskeleton robot," *World J. Eng. Technol.*, vol. 7, no. 1, pp. 68–79, 2019. [Online]. Available: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/wjet.2019.71004

[4] Z. Jia-Yong, L. Ye, M. Xin-Min, H. Chong-Wei, M. Xiao-Jing, L. Qiang, W. Yue-Jin, and Z. Ang, "A preliminary study of the military applications and future of individual exoskeletons," *J. Phys., Conf. Ser.*, vol. 1507, no. 10, Mar. 2020, Art. no. 102044. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1507/10/102044

[5] S. Fox, O. Aranko, J. Heilala, and P. Vahala, "Exoskeletons: Comprehensive, comparative and critical analyses of their potential to improve manufacturing performance," *J. Manuf. Technol. Manage.*, vol. 31, no. 6, pp. 1261–1280, Jun. 2020. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JMTM-01-2019-0023/full/html

[6] N. Aliman, R. Ramli, and S. M. Haris, "Design and development of lower limb exoskeletons: A survey," *Robot. Auto. Syst.*, vol. 95, pp. 102–116, Sep. 2017.

[7] G. Bao, L. Pan, H. Fang, X. Wu, H. Yu, S. Cai, B. Yu, and Y. Wan, "Academic review and perspectives on robotic exoskeletons," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2294–2304, Nov. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8853289/

[8] H. Kazerooni and R. Steger, "The Berkeley lower extremity exoskeleton," *J. Dyn. Syst., Meas., Control*, vol. 128, no. 1, pp. 14–25, Mar. 2006. [Online]. Available: https://asmedigitalcollection.asme.org/dynamicsystems/article/128/1/14/465257/The-Berkeley-Lower-Extremity-Exoskeleton

[9] (Oct. 26, 2020). *Human Universal Load Carrier (HULC)*. [Online]. Available: https://www.army-technology.com/projects/human-universal-load-carrier-hulc/

[10] S. Karlin, "Raiding iron man's closet [geek life]," *IEEE Spectr.*, vol. 48, no. 8, p. 25, Aug. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5960158/

[11] Y. Sankai, "HAL: Hybrid assistive limb based on cybernics," in *Robotics Research* (Springer Tracts in Advanced Robotics), vol. 66. Berlin, Germany: Springer, 2010, pp. 25–34. [Online]. Available: http://link.springer.com/10.1007/978-3-642-14743-2_3

[12] M. Fontana, R. Vertechy, S. Marcheschi, F. Salsedo, and M. Bergamasco, "The body extender: A full-body exoskeleton for the transport and handling of heavy loads," *IEEE Robot. Autom. Mag.*, vol. 21, no. 4, pp. 34–44, Dec. 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6990863/

[13] Y. Ma, X. Wu, J. Yi, C. Wang, and C. Chen, "A review on human-exoskeleton coordination towards lower limb robotic exoskeleton systems," *Int. J. Robot. Autom.*, vol. 34, no. 4, pp. 431–451, 2019. [Online]. Available: http://www.actapress.com/PaperInfo.aspx?paperId=46280

[14] T. Yan, M. Cempini, C. M. Oddo, and N. Vitiello, "Review of assistive strategies in powered lower-limb orthoses and exoskeletons," *Robot. Auto. Syst.*, vol. 64, pp. 120–136, Feb. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889014002176

[15] H. F. N. Al-Shuka, M. H. Rahman, S. Leonhardt, I. Ciobanu, and M. Berteanu, "Biomechanics, actuation, and multi-level control strategies of power-augmentation lower extremity exoskeletons: An overview," *Int. J. Dyn. Control*, vol. 7, no. 4, pp. 1462–1488, Dec. 2019, doi: 10.1007/s40435-019-00517-w.

[16] H. F. N. Al-Shuka, R. Song, and C. Ding, "On high-level control of power-augmentation lower extremity exoskeletons: Human walking intention," in *Proc. 10th Int. Conf. Adv. Comput. Intell. (ICACI)*, Mar. 2018, pp. 169–174. [Online]. Available: https://ieeexplore.ieee.org/document/8377601/

[17] J. Taborri, E. Palermo, S. Rossi, and P. Cappa, "Gait partitioning methods: A systematic review," *Sensors*, vol. 16, no. 1, p. 66, Jan. 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/1/66

[18] H. F. N. Al-Shuka and R. Song, "On low-level control strategies of lower extremity exoskeletons with power augmentation," in *Proc. 10th Int. Conf. Adv. Comput. Intell. (ICACI)*, Mar. 2018, pp. 63–68. [Online]. Available: https://ieeexplore.ieee.org/document/8377581/

[19] W. Huo, S. Mohammed, J. C. Moreno, and Y. Amirat, "Lower limb wearable robots for assistance and rehabilitation: A state of the art," *IEEE Syst. J.*, vol. 10, no. 3, pp. 1068–1081, Sep. 2016. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6930719

[20] H. Kazerooni, J.-L. Racine, L. Huang, and R. Steger, "On the control of the Berkeley lower extremity exoskeleton (BLEEX)," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2005, pp. 4353–4360. [Online]. Available: http://ieeexplore.ieee.org/document/1570790/

[21] H. Kim, Y. J. Shin, and J. Kim, "Design and locomotion control of a hydraulic lower extremity exoskeleton for mobility augmentation," *Mechatronics*, vol. 46, pp. 32–45, Oct. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957415817300880

[22] J. Ghan, R. Steger, and H. Kazerooni, "Control and system identification for the Berkeley lower extremity exoskeleton (BLEEX)," *Adv. Robot.*, vol. 20, no. 9, pp. 989–1014, Jan. 2006. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1163/156855306778394012

[23] T. Hayashi, H. Kawamoto, and Y. Sankai, "Control method of robot suit HAL working as operator's muscle using biological and dynamical information," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 2, no. 1, Aug. 2005, pp. 3063–3068. [Online]. Available: http://ieeexplore.ieee.org/document/1545505/

[24] S. Lee and Y. Sankai, "Virtual impedance adjustment in unconstrained motion for an exoskeletal robot assisting the lower limb," *Adv. Robot.*, vol. 19, no. 7, pp. 773–795, Jan. 2005. [Online]. Available: https://www.tandfonline.com/doi/full/10.1163/1568553054455095

[25] S. N. Yu, H. D. Lee, S. H. Lee, W. S. Kim, J. S. Han, and C. S. Han, "Design of an under-actuated exoskeleton system for walking assist while load carrying," *Adv. Robot.*, vol. 26, nos. 5–6, pp. 561–580, Jan. 2012. [Online]. Available: https://www.tandfonline.com/doi/full/10.1163/156855311X617506

[26] W. S. Kim, H. D. Lee, D. H. Lim, C. S. Han, and J. S. Han, "Development of a lower extremity exoskeleton system for walking assistance while load carrying," in *Nature-Inspired Mobile Robotics*. Singapore: World Scientific, Aug. 2013, pp. 35–42. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/9789814525534_0008

[27] W. S. Kim, H. D. Lee, D. H. Lim, J. S. Han, K. S. Shin, and C. S. Han, "Development of a muscle circumference sensor to estimate torque of the human elbow joint," *Sens. Actuators A, Phys.*, vol. 208, pp. 95–103, Feb. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924424713006341

[28] S. Yu, H. Lee, W. Kim, and C. Han, "Development of an underactuated exoskeleton for effective walking and load-carrying assist," *Adv. Robot.*, vol. 30, no. 8, pp. 535–551, Apr. 2016. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/01691864.2015.1135080

[29] H. Kazerooni, R. Steger, and L. Huang, "Hybrid control of the Berkeley lower extremity exoskeleton (BLEEX)," *Int. J. Robot. Res.*, vol. 25, nos. 5–6, pp. 561–573, May 2006. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364906065505

[30] K. Low, X. Liu, and H. Yu, "Development of NTU wearable exoskeleton system for assistive technologies," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, vol. 2, Jul. 2005, pp. 1099–1106. [Online]. Available: http://ieeexplore.ieee.org/document/1626705/

[31] K. H. Low, X. Liu, C. H. Goh, and H. Yu, "Locomotive control of a wearable lower exoskeleton for walking enhancement," *J. Vibrat. Control*, vol. 12, no. 12, pp. 1311–1336, Dec. 2006. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1077546306070616

[32] D. Lim, W. Kim, H. Lee, H. Kim, K. Shin, T. Park, J. Lee, and C. Han, "Development of a lower extremity exoskeleton robot with a quasi-anthropomorphic design approach for load carriage," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 5345–5350. [Online]. Available: http://ieeexplore.ieee.org/document/7354132/

[33] C. Zhang, X. Zang, Z. Leng, H. Yu, J. Zhao, and Y. Zhu, "Human–machine force interaction design and control for the HIT load-carrying exoskeleton," *Adv. Mech. Eng.*, vol. 8, no. 4, pp. 1–14, Apr. 2016. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1687814016645068

[34] S. Chen, Z. Chen, and B. Yao, "Precision cascade force control of multi-DOF hydraulic leg exoskeleton," *IEEE Access*, vol. 6, pp. 8574–8583, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8279424/

[35] C.-F. Chen, Z.-J. Du, L. He, J.-Q. Wang, D.-M. Wu, and W. Dong, "Active disturbance rejection with fast terminal sliding mode control for a lower limb exoskeleton in swing phase," *IEEE Access*, vol. 7, pp. 72343–72357, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8721059/

[36] C.-F. Chen, Z.-J. Du, L. He, Y.-J. Shi, J.-Q. Wang, G.-Q. Xu, Y. Zhang, D.-M. Wu, and W. Dong, "Development and hybrid control of an electrically actuated lower limb exoskeleton for motion assistance," *IEEE Access*, vol. 7, pp. 169107–169122, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8897544/

[37] J. Jiang, Y. Wang, H. Cao, J. Zhu, W. Zhu, and L. Jin, "On the control of lower extremity exoskeleton base on the interaction force of torso," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2020, pp. 90–95.

[38] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8103164/

[39] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Found. Trends Mach. Learn.*, vol. 11, nos. 3–4, pp. 219–354, 2018. [Online]. Available: http://www.nowpublishers.com/article/Details/MAL-071

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[41] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "DeepMimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Jul. 2018, doi: 10.1145/3197517.3201311.

[42] N. Chentanez, M. Müller, M. Macklin, V. Makoviychuk, and S. Jeschke, "Physics-based motion capture imitation with deep reinforcement learning," in *Proc. 11th Annu. Int. Conf. Motion, Interact., Games*. New York, NY, USA: ACM, Nov. 2018, pp. 1–10. [Online]. Available: https://dl.acm.org/doi/10.1145/3274247.3274506

[43] W. Yu, G. Turk, and C. K. Liu, "Learning symmetric and low-energy locomotion," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–12, Aug. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3197517.3201397

[44] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: A comprehensive survey," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 945–990, Feb. 2022. [Online]. Available: https://link.springer.com/10.1007/s10462-021-09997-9

[45] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Sci. Robot.*, vol. 4, no. 26, p. 5872, Jan. 2019.

[46] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Sci. Robot.*, vol. 5, no. 47, Oct. 2020, Art. no. eabc5986. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abc5986

[47] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Sci. Robot.*, vol. 7, no. 62, p. 2822, Jan. 2022. [Online]. Available: https://www.science.org/doi/10.1126/scirobotics.abk2822

[48] X. Bin Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," in *Robotics: Science and Systems XVI*. Robotics: Science and Systems Foundation, Jul. 2020. [Online]. Available: http://www.roboticsproceedings.org/rss16/p064.pdf

[49] S. Ha, P. Xu, Z. Tan, S. Levine, and J. Tan, "Learning to walk in the real world with minimal human effort," in *Proc. 4th Conf. Robot Learn. (CoRL)*, Cambridge, U.K., 2020, pp. 1–11.

[50] Z. Xie, G. Berseth, P. Clary, J. Hurst, and M. van de Panne, "Feedback control for Cassie with deep reinforcement learning," in *Proc. EEE Int. Conf. Intell. Robots Syst.*, Oct. 2018, pp. 1241–1246.

[51] 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 *Proc. Conf. Robotic Learn.*, 2019, pp. 1–13.

[52] F. Abdolhosseini, H. Y. Ling, Z. Xie, X. B. Peng, and M. van de Panne, "On learning symmetric locomotion," in *Motion, Interaction and Games*. New York, NY, USA: ACM, Oct. 2019, pp. 1–10. [Online]. Available: https://dl.acm.org/doi/10.1145/3359566.3360070

[53] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for robust parameterized locomotion control of bipedal robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 2811–2817. [Online]. Available: https://ieeexplore.ieee.org/document/9560769/

[54] J. Ahn, J. Lee, and L. Sentis, "Data-efficient and safe learning for humanoid locomotion aided by a dynamic balancing model," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4376–4383, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9079565/

[55] L. C. Melo and M. R. O. A. Máximo, "Learning humanoid robot running skills through proximal policy optimization," in *Proc. Latin Amer. Robot. Symp. (LARS), Brazilian Symp. Robot. (SBR) Workshop Robot. Educ. (WRE)*, Oct. 2019, pp. 37–42. [Online]. Available: https://ieeexplore.ieee.org/document/9018554/

[56] C. Yang, T. Komura, and Z. Li, "Emergence of human-comparable balancing behaviours by deep reinforcement learning," in *Proc. IEEE-RAS 17th Int. Conf. Humanoid Robot. (Humanoids)*, Nov. 2017, pp. 372–377. [Online]. Available: http://ieeexplore.ieee.org/document/8246900/

[57] C. Yang, K. Yuan, S. Heng, T. Komura, and Z. Li, "Learning natural locomotion behaviors for humanoid robots using human bias," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2610–2617, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8990011/

[58] R. Özaln, C. Kaymak, Ö. Yildirum, A. Ucar, Y. Demir, and C. Güzelis, "An implementation of vision based deep reinforcement learning for humanoid robot locomotion," in *Proc. IEEE Int. Symp. Innov. Intell. Syst. Appl. (INISTA)*, Jul. 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8778209/

[59] R. Zheng, Z. Yu, H. Liu, Z. Zhao, J. Chen, and L. Jia, "Sensitivity adaptation of lower-limb exoskeleton for human performance augmentation based on deep reinforcement learning," *IEEE Access*, vol. 11, pp. 36029–36040, 2023.

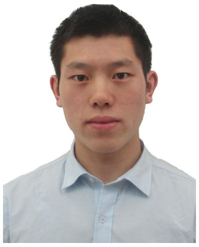
[60] X. B. Peng and M. van de Panne, "Learning locomotion skills using DeepRL: Does the choice of action space matter?" in *Proc. ACM SIGGRAPH/Eurographics Symp. Comput. Animation (SCA)*, no. 1. New York, NY, USA: ACM Press, 2017, pp. 1–13. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3099564.3099567

[61] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 4, Feb. 2018, pp. 2587–2601.

[62] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.

[63] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: http://www.nature.com/articles/nature14236

[64] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "GBR DeepMind 2014 deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, vol. 1, 2014, pp. 605–619.

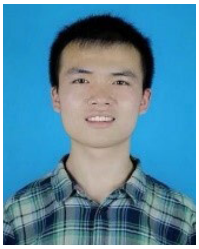


RANRAN ZHENG received the B.S. degree in flight vehicle propulsion engineering from the Beijing Institute of Technology, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Flight Vehicle Control, School of Aerospace Engineering. His main research interests include mechanical design, modeling and control of exoskeletons, and locomotion control based on deep reinforcement learning.

ZHIYUAN YU received the Ph.D. degree in aircraft control from the Beijing Institute of Technology, Beijing, China, in 2009. He is currently with the Laboratory of Aerospace Servo Actuation and Transmission, Beijing Institute of Precision Mechatronics and Controls. His current research interests include aerospace servo actuation and transmission, wearable robots, and servo motor.

HONGWEI LIU received the B.S. and M.S. degrees in exploration guidance and control engineering from the Department of Flight Vehicle Control, School of Aerospace Engineering, Beijing Institute of Technology, in 2014 and 2017, respectively. He is currently with the Laboratory of Aerospace Servo Actuation and Transmission, Beijing Institute of Precision Mechatronics and Controls. His research interests include signal processing, system identification, and control of wearable robotic devices.

JING CHEN received the Ph.D. degree from the China Academy of Launch Vehicle Technology, Beijing, China, in 2020. She is currently with the Laboratory of Aerospace Servo Actuation and Transmission, Beijing Institute of Precision Mechatronics and Controls. Her current research interests include aerospace servo actuation and transmission, and motor and exoskeleton robot control.



ZHE ZHAO received the M.S. degree in mechanical engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2019. He is currently with the Laboratory of Aerospace Servo Actuation and Transmission, Beijing Institute of Precision Mechatronics and Controls. His current research interests include mechanism design of exoskeleton and intelligent mechanism design.

LONGFEI JIA received the Ph.D. degree from the Beijing Institute of Precision Mechatronics and Controls, Beijing, China, in 2022. He is currently with the Laboratory of Aerospace Servo Actuation and Transmission, Beijing Institute of Precision Mechatronics and Controls. His research interests include kinematics, dynamics, and intelligence control of robots.