Journal of Institute of Control, Robotics and Systems (2025) 31(2):106-113
https://doi.org/10.5302/J.ICROS.2025.24.0270 ISSN:1976-5622 eISSN:2233-4335

험난한 지형에서의 강건한 사족보행 제어를 위한 대칭적 파쿠르 학습

Symmetric Parkour Learning for Robust Quadruped Robot Control on Challenging Terrain

김 동 주, 박 지 훈, 이 인 호*
(Dong-Ju Kim[1], Jihoon Park[1], and Inho Lee[1,*])
[1]Department of Electrical and Electronics Engineering, Pusan National University

Abstract: Parkour tasks on various terrains are challenging for quadruped robots, requiring high levels of adaptability and robustness. This paper presents a novel approach to enable quadruped robots to perform parkour tasks on challenging terrain through symmetric reinforcement learning. We propose a modified PPO (Proximal Policy Optimization) that leverages the robot's symmetry loss to enhance its stability and improve the efficiency of training convergence. This symmetric learning method brings generalization, allowing the robot to be robust even in previously unseen environments. The effectiveness of our approach is demonstrated through results showing both learning convergence efficiency and robust performance on new, challenging terrains when applying the trained model.

Keywords: Quadruped Robot, reinforcement learning, symmetry

I. 서론

로봇의 학습 기반 보행 기술이 발전하면서 점차 복잡한 환경에서 다양한 동작을 수행할 수 있는 기술이 개발되고 있다. 초기에는 카메라나 라이더 같은 외부센서 없이 로봇의 내부센서인 IMU나 각 조인트의 모터 엔코더를 통한 보행에 초점이 맞춰졌다[1-4]. 하지만 이러한 고유 감각정보만을 가지고 보행을 하게 되면 불규칙적인 지형이나 험난한 지형 및 어려운 태스크에서는 한계가 있다. 이에 따라 최근에는 외부센서를 결합한 형태로 발전했고[5], 보다 고도화된 보행과 장애물 회피가 가능해졌다. 이러한 기술을 확장하여 파쿠르와 같은 고난도 동작을 수행할 수 있도록 학습기반 로봇 보행 연구가 진행되고 있다[6-8]. 파쿠르는 장애물을 넘나들고 회피하는 다양한 움직임을 요구하여, 보다 적응성 높은 행동을 수행하여야 한다. 외부센서를 사용하지 않는 블라인드 보행 제어와 비교하여, 높은 보행 수준을 달성하기 위해서는 추가 적인 정보를 제공하여야 한다. 예를 들어 로봇과 주변 지형의 높이 차이 정보 혹은 깊이 정보를 이용할 수 있다.

여기서는 사족보행 로봇의 보행 적응력을 한 단계 더 발전시키기 위해 대칭 학습법을 제안한다. 대칭성은 속도와 에너지 효율성에서 이점을 제공하여, 로봇이 다양한 상황에서 유사한 패턴으로 대응할 수 있게 한다[9]. 강화학습을 통한 대칭적 정보 학습을 통해 로봇은 학습 중 경험하지 못한 불확실한 환경에서도 강건한 보행이 가능하며, 예기치 않은 충격이나 변형된 지형에서도 안정성을 유지할 수 있다[10-11].

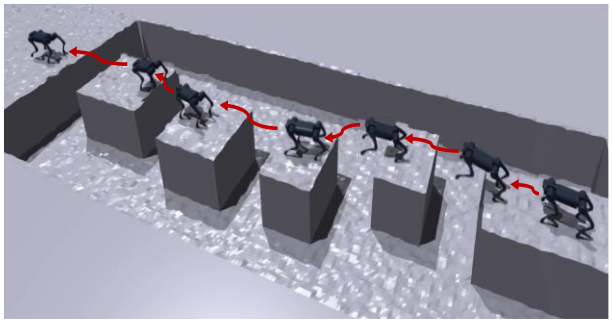


그림 1. 대칭 학습에 의한 시뮬레이션 Isaac Gym에서 진행된 험지 보행.
Fig. 1. Rough terrain locomotion in Isaac Gym simulation with proposed symmetric learning.

본 연구에서는 파쿠르와 같은 도전적인 환경에서 안정적인 보행을 목표로 하는 강화학습 기반 대칭 학습 프레임워크를 제안한다. 기존 PPO 강화학습 알고리즘에 대칭성을 고려한 손실 함수를 추가하여 학습 수렴 속도를 높이고, 학습 중 경험해보지 못한 미지의 환경에서도 민첩하게 장애물을 넘으며 이동할 수 있도록 한다. 또한 외란이 작용하는 환경

* Corresponding Author
Manuscript received November 10, 2024; revised December 4, 2024; accepted December 28, 2024
김동주: 부산대학교 전기전자공학과 대학원생(ehdwn70001@pusan.ac.kr, ORCID 0009-0008-5695-3325)
박지훈: 부산대학교 전기전자공학과 대학원생(kyo2z5jk@pusan.ac.kr, ORCID 0000-0001-6061-3202)
이인호: 부산대학교 전기전자공학과 교수(inholee8@pusan.ac.kr, ORCID 0000-0002-5046-5207)
※ 본 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2023-00215760, 가이드 독: 시각장애인 길 안내 로봇 이동지능 기술 개발, 50%)과 2024년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(RS-2024-00406796, 2024년 산업혁신인재성장지원사업, 50%).

에서도 강건한 보행을 할 수 있도록 한다.

II. 학습 전략 및 파쿠르 시스템

본 논문에서는 Unitree사의 사족보행 A1를 시뮬레이션 환경 Isaac Gym에서 학습하였다(그림 1)[12]. 본 내용에서는 대칭 학습 방법에 대한 소개를 한다. 또한 강화학습 환경 구성 및 시스템 개요에 대해서도 설명을 한다.

1. 대칭적 학습 방법론

대칭성을 강화학습의 프레임워크에 적용함으로써, 로봇이 경험하지 못한 불확실한 환경에서 강건한 보행을 할 수 있다. 본 장에서는 현재 상태의 대칭성을 고려한 MDP (Markov Decision Process) 변환을 설명한다. 또한 대칭적 학습법은 기존의 PPO가 업데이트되는 손실 함수에 대칭 손실 함수가 추가되어 업데이트된다. 따라서 간단하게 PPO 알고리즘에 대해서 설명하고 MSL (Mirror Symmetry Loss), PSL (Proximal Symmetry Loss) [10]과 이들의 적용에 대해서도 내용을 전달하고자 한다.

1.1. 기본 개념

강화학습은 MDP로 표현 가능하다. MDP는 $\langle S, A, \psi, p, r\rangle$로 구성되는데, S는 에이전트의 상태의 집합, A는 에이전트가 취하는 행동의 집합, ψ는 가능한 상태, 행동의 쌍, p는 특정 상태 s에서 행동 a를 하였을 때 다음 상태 s'로 전이될 확률의 집합이고 마지막 r은 에이전트가 상태 s에서 행동 a를 취했을 때 받는 보상의 집합이다. 대칭성을 활용한 학습에서는 MDP의 상태와 행동을 변환하여 새로운 상태-행동 쌍을 구성한다. 특정 상태에서 행동이 대칭 상태에서도 동일하게 작동하도록 학습하여 정책의 일관성을 유지한다. 이를 위해서는 MDP mapping을 통해 상태, 행동 쌍에 대해서도 각각 변환해주어야 한다. 기존의 상태의 s를 대칭적 상태 $f(s)$로 변환하고 행동 a를 행동 변환 $g_s(a)$로 변환한다. 본 연구에서는 로봇의 진행 방향 (x축)의 좌우 대칭 변환을 하여 로봇의 상태 값인 조인트의 각도, 속도 및 행동 a를 진행 방향 기준으로 대칭 변환시켰다. 이러한 변환을 통해 대칭 상태와 행동을 구성할 수 있으며, 이를 만족시키는 정책은 다음 식을 만족한다.

$$p\big(f(s), g_s(a), f(s')\big) = p(s, a, s'), \forall s, s' \in S, \forall a \in A \quad (1)$$

$$r\big(f(s), g_s(a)\big) = r(s, a), \ \forall s \in S, \forall a \in A \quad (2)$$

(1), (2)식은 각각 상태-행동 쌍 (s, a)와 그 대칭 쌍 $\big(f(s), g_s(a)\big)$가 같은 전이확률 및 보상 값을 가진다는 것을 의미하고, 이것은 로봇이 훈련 중에 경험하지 못한 상태와 행동에 대해서도 학습이 가능하다는 장점이 있다.

1.2 PPO (Proximal Policy Optimization)

PPO 알고리즘은 강화학습에서 안정적인 정책 학습을 위해 고안된 손실 함수를 사용한다[13]. 정책이 크게 업데이트 되는 것을 방지하기 위해 클램핑(clamping)을 적용하여, 안전 범위를 제한한다. 또한 가치 함수 손실과 탐색을 촉진하기 위한 엔트로피 항을 포함하여 학습의 안정과 효율성을 높일 수 있다. 이 3가지 항에 대한 식은 다음과 같이 나타낼 수 있다.

$$L^{Clip}(\theta) = E_t\big[\min\big(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\big] \quad (3)$$

$$with \quad r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}, \ \hat{A}_t = Q(s_t, a_t) - V(s_t) \quad (4)$$

$$L^{\mathrm{VF}}(\omega) = E_t\big[V_\omega(s_t) - V_t^{\mathrm{target}}\big]^2 \quad (5)$$

$$H(\pi_\theta) = -E_t[\textstyle\sum_a \pi_\theta(a_t|s_t)\log\pi_\theta(a_t|s_t)] \quad (6)$$

$$L^{PPO}(\theta, \omega) = L^{Clip}(\theta) - c_1 L^{VF}(\omega) + c_2 H(\pi_\theta) \quad (7)$$

식 (3), (4)는 정책 개선을 유도하면서 과도한 업데이트를 방지하는 surrogate loss 값이고, 여기서 정책 $\pi_\theta(a_t|s_t)$는 상태 s_t에서 행동 a_t를 선택할 확률 분포를 나타내며, 이 확률 분포는 매개변수 θ 에 의해 결정된다. $r_t(\theta)$는 과거정책과 새로운 정책의 비율을 매개변수 ϵ을 통해 조절한다. $\hat{A}_t$는 어드밴티지 함수로 특정 상태 s_t, 행동 a_t의 가치 $Q(s_t, a_t)$가 평균적인 상태에서의 가치 $V(s_t)$보다 얼마나 더 나은지를 나타낸다. PPO에서는 일반적으로 GAE (Generalized Advantage Estimation) 방법을 통해 어드밴티지 $\hat{A}_t$를 추정한다[13]. 식 (5)는 가치 손실 함수로, 상태 가치 $V_\omega(s_t)$와 실제 목표 가치 V_t^{target} 간의 차이를 최소화하는 역할을 하여 정책의 가치를 평가하는 크리틱(critic)이 매개변수 ω에 의해 업데이트되어, 상태의 가치를 정확히 평가할 수 있게 한다. 식 (6)은 정책의 불확실성을 증가시키는 역할을 하여, 에이전트의 탐색 (exploration)을 확대시켜 다양한 행동을 시도하도록 한다. 따라서 식 (7)의 PPO 손실함수가 도출되고 이 값을 최대화 하는 방향으로 정책이 학습된다.

1.3 대칭 학습

지금부터는 전 장에 설명했던, PPO에 대칭을 고려한 강화학습 알고리즘인 symmetric learning에 대해서 설명할 것이다. 이 장에서 설명하는 손실 함수는 대칭 손실 함수를 의미한다. 본 논문에서는 2가지 방법을 통해 기존의 PPO의 결과와 비교하려고 한다. 첫 번째 방법은 MSL, 두 번째 방법은 PSL이다[10-11]. 두 방법의 차이는 액터 손실 함수 $L_t^\pi(\theta)$를 구하는 데 있다.

우선 MDP 대칭 변환으로부터 구한 상태-행동 쌍으로부터 대칭 손실 함수(symmetric loss function)를 구할 수 있다.

$$L^S(\theta, \omega) = E_t[w_\pi \cdot L_t^\pi(\theta) + w_V \cdot L_t^V(\omega)] \quad (8)$$

$$L_t^{\pi,MSL}(\theta) = E_t\big\|\pi_\theta(s_t) - g_{f(s_t)}(\pi_\theta(f(s_t)))\big\|^2 \quad (9)$$

$$L_t^V(\omega) = E_t\left[\big(V_\omega\big(f(s_t)\big) - V_t^{\mathrm{targ}}\big)^2\right] \quad (10)$$

$$L^{\pi,\mathrm{PSL}}(\theta) = -E_t[\min(x_t(\theta), 1+\epsilon)] \quad (11)$$

$$with \quad x_t(\theta) = \frac{\min\Big(\pi_\theta\big(\bar{a}_t|f(s_t)\big), \pi_{\theta_{\mathrm{old}}}(a_t|s_t)\Big)}{\pi_{\theta_{\mathrm{old}}}\big(\bar{a}_t|f(s_t)\big)} \quad (12)$$

$$and \ \bar{a}_t = g_s(\pi_{\theta_{\mathrm{old}}}(s_t \mid \sigma = 0)) \quad (13)$$

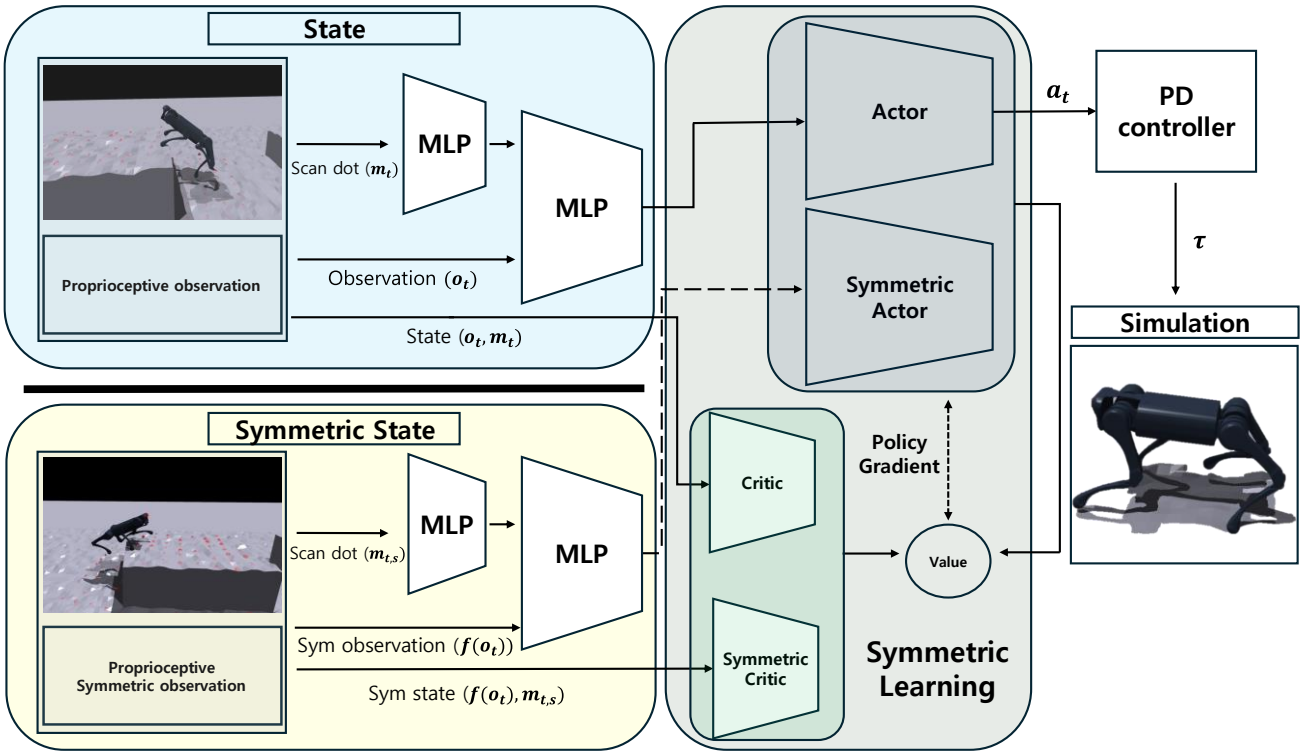


그림 2. 대칭적 강화학습 프레임워크.

Fig. 2. The framework of symmetric reinforcement learning.

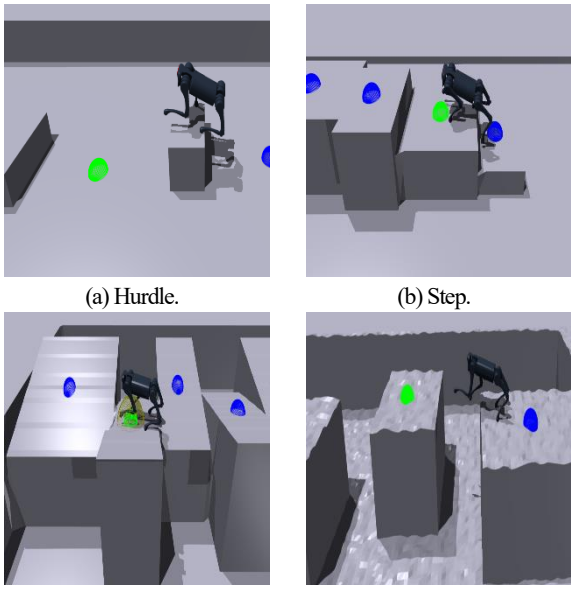
식 (8)은 대칭 손실 함수를 나타낸 것으로 w_π, w_V 는 각각 정책의 액터(actor)와 크리틱(critic) 대칭 손실 함수의 계수 값이다. $L_t^\pi(\theta)$, $L_t^V(\omega)$는 액터와 크리틱의 대칭 손실 함수로, θ, ω은 각각 액터, 크리틱의 매개변수로 매 스텝마다 업데이트 되는 파라미터 값이다. 식 (9)를 보면, mirror symmetric의 액터 손실 함수는 상태-행동 쌍 (s_t, a_t)으로부터 구한 정책의 값 $\pi_\theta(s_t)$과 상태-행동쌍 $(f(s), g_s(a))$으로부터 구한 대칭 변환 정책 $g_{f(s_t)}(\pi_\theta(f(s_t)))$의 MSE (Mean Square Error)로 구할 수 있다. 크리틱 손실 함수 $L_t^V(\omega)$ 는 식 (10)을 통해 타겟 가치와 대칭 변환된 가치 함수의 MSE로 구한다.

반면 두 번째 방법인 proximal symmetric은 PPO의 신뢰 구간 개념을 기반으로 하여, 정책이 대칭성을 유지하거나 점진적으로 개선될 수 있도록 학습된다[11]. 식 (11)~(13)을 보면, PSL의 핵심인 대칭 확률 비율 $x_t(\theta)$ 는 이전 정책과 현재 정책의 비율로 정의된다. $\pi_{\theta_{\text{old}}}(\cdot \mid s_t)$ 는 업데이트 이전의 정책을 나타내며, 확률적 행동(stochastic action)이 아닌 결정론적 행동(deterministic action)이다. $\pi_{\theta_{\text{old}}}(s_t \mid \sigma = 0)$ 는 이전 시점 정책 분포의 표준편차 $\sigma = 0$ 일 때의 결정론적 행동이다. 따라서 현재 시점의 정책 업데이트에서 상수가 아닌 값은 오직 $\pi_\theta(\bar{a}_t \mid f(s_t))$이며, 최적화되는 매개변수는 θ 뿐이다. 대칭 상태 $f(s_t)$는 실제로 경험하지 못한 가상의 대칭 변환 상태이며, 이는 탐색 상태 s_t 를 변환한 값이다. 이러한 방법을 통해 PSL은 기존의 탐색 상태 s_t 에 대한 정책을 유지하면서, 대칭 상태에서의 정책 $\pi_\theta(f(s_t))$ 만을 조정한다. 또한 이 과정은 클램핑 기법을 사용하여 과도한 업데이트를 방지함으로써 학습의 안정성을 확보할 수 있다.

MSL은 식 (8)에서 나타나듯이, 탐색 상태 s_t와 대칭 상태 $f(s_t)$ 모두에서 정책 π_θ를 동시에 조정하려고 한다. 이 방식은 대부분의 경우에는 적합하지만, 탐색 불균형이 발생하면 문제가 생긴다. 예를 들어 왼쪽 장애물과 오른쪽 장애물을 넘는 경로를 학습하는 상황을 가정해보자. 로봇이 왼쪽으로 경로를 더 많이 탐색한 경우, 왼쪽의 경우에는 충분히 탐색해서 이미 최적화되었을 가능성이 높다. 반면, 오른쪽 경로는 탐색이 부족하여 정책 확률 분포가 최적화 되지 않았을 수 있다. 여기서 문제가 발생하는데, 식 (9)를 살펴보면 MSL방법은 탐색 상태 s_t 와 대칭 상태 $f(s_t)$ 의 정책을 모두 업데이트하려고 하기 때문에, 왼쪽 경로에서 잘 학습된 정책이 오른쪽 경로로 인해 잘못된 방향으로 업데이트 될 가능성이 있다. 이로 인해 이미 최적화된 정책 네트워크가 망가질 수 있다. 반면 PSL은 앞서 언급하였듯이, 대칭 상태 $f(s_t)$ 에서의 정책 $\pi_\theta(f(s_t))$ 만을 조정하며, 탐색 상태에 대한 정책은 그대로 유지된다. 이는 탐색 상태에서의 정책 네트워크를 보호하면서도, 대칭 상태에서의 행동이 탐색 상태와의 일관성을 유지하도록 한다. 결론적으로 로봇이 다양한 방향의 지형에서도 안정적으로 보행할 수 있도록 한다. 기존의 PPO에 대칭 학습법을 고려하게 되면 식 (14)와 같은 형태로 같은 최종 목적 함수가 결정된다.

$$L^{PPO+S}(\theta, \omega) = L^{PPO}(\theta, \omega) - L^S(\theta, \omega) \qquad (14)$$

본 연구에서는 대칭 상태 $f(s)$ 를 로봇의 진행방향 축(x축)을 기준으로 구성하였고, 두 가지 대칭 학습 기법을 모두 적용하여 기존보다 빠른 수렴속도, 높은 보상함수 값, 훈련 중에 경험하지 못한 상태에서의 강건함을 확인하였다.



(a) Hurdle. (b) Step.

(c) Tilted Ramp. (d) Gap.

그림 3. 학습에 사용된 네 가지 지형 유형.

Fig. 3. Four types of terrain used in training.

2. 파쿠르 학습

일반적으로 사족보행 로봇이 장애물이 있는 파쿠르 태스크를 수행하기는 쉽지 않다[8]. 이를 성공적으로 수행하기 위해서는 다음 장애물을 넘기 위한 움직임을 학습할 수 있는 충분한 관측값과 보상함수의 설계가 중요하다. 본 장에서는 사족보행 A1을 대상으로 한 파쿠르 학습 환경을 설명한다. 파쿠르 태스크에 사용된 관측값과, 보상 함수의 디자인에 대해 다루며, 앞서 소개한 대칭 학습 개념을 파쿠르 태스크에 적용한 방법론에 대해서도 얘기를 한다.

그림 (2)를 보면, 파쿠르 대칭 학습의 프레임워크를 알 수 있다. 행동 정책 네트워크를 생성하는 액터에서는 IMU, 엔코더로부터 구할 수 있는 조인트 각도, 속도 등의 고유 감각 정보를 통해 가치가 높아지는 방향으로 업데이트한다. 여기서 가치를 평가하는 크리틱으로부터 정책 그래디언트가 일어나는데, 본 연구에서는 PPO 알고리즘에 대칭 상태-액션쌍을 고려한 대칭 액터, 크리틱의 정보를 추가로 고려하여 정책이 업데이트 된다. 여기서 대칭 액터로부터 출력되는 행동은 실제 로봇의 컨트롤러에 인풋으로 들어 가지 않고 단지 정책을 평가하는 데에만 쓰인다. 이러한 정책 업데이트를 통해 실제 로봇의 컨트롤러에 인풋으로 들어가는 액터의 행동쌍들이 PD제어를 통해 토크를 생성해 로봇의 움직임을 만들어낸다. 여기서, 파쿠르 태스크는 그림 (3)과 같은 다양한 험지 및 어려운 환경에서 훈련이 되었다.

2.1 파쿠르 태스크

로봇의 상태 값부터 자세하게 살펴보자면, 식 (15)처럼 관측값 o_t가 정의된다

$$o_t = [\omega_t, \theta, \psi, h_t, q_t, \dot{q}_t, a_{t-1}, c, v_t] \quad (15)$$

ω_t 는 로봇프레임 기준의 각속도이고, θ, ψ 는 로봇의 roll, pitch 각도를 나타낸다. 그 다음에 나오는 h_t 는 현재 로봇과 다음 목표 지점까지의 yaw방향 각도, $q_t, \dot{q}_t$ 는 각각 조인트의 각도와 속도를 의미한다. a_{t-1} 는 이전 상태에서의 액션 값이고, c는 로봇 컨택 여부를 0,1의 이진수로 나타낸 것이다. v_t 는 로봇의 선속도를 의미하고 여기서는 이 값을 구할 때, MLP (Multi-Layer Perceptron)로부터 속도를 구하였다[14]. 로봇의 관측값은 일반적으로 선속도 v_t, 각속도 ω_t, 몸의 회전에 해당하는 imu 값, 모터 엔코더로부터 구할 수 있는 조인트 위치 q_t, 속도 $\dot{q}_t$ 그리고 이전 스텝에서의 로봇 액션 a_{t-1} 을 쓴다[14]. 하지만 파쿠르의 경우에는 도전적인 (challenging) 태스크이므로 추가적인 정보가 필요하다. 여기서는 로봇의 컨택여부를 추가적인 관측값으로 사용하였다. 또한 로봇의 상태 값이 다음 행동을 하기에 충분한 정보를 제공해야 하기 때문에 주변 지형에 대한 높이 정보를 이용하였다[7].

그림 (2)를 보면 알 수 있듯이, 주변 지형 높이인 m_t 와 로봇의 관측 값 o_t를 결합한 상태 값을 입력으로 받아 액터, 크리틱 네트워크를 통과한다. 최종적으로 액터 네트워크를 거쳐서 로봇의 액션인 a_t 를 구한다. 이 액터 네트워크는 12차원의 액션을 출력해내고, 이 액션값은 조인트의 타겟 (target) 위치 값을 구하는 데 정의된다.

$$q_{des} = q_{nominal} + \alpha * a_t \quad (16)$$

식 (16)을 보면, $q_{nominal}$ 은 로봇이 서 있는 초기 위치의 조인트 값이고 α 는 액션에 곱해지는 상수 값이다. q_{des}로부터 PD제어를 통해 토크 값 τ을 구할 수 있다.

$$\tau = K_p(q_{des} - q) - K_d(\dot{q}) \quad (17)$$

K_p = 40, K_d= 1.0을 사용하였다. 이렇게 최종 토크 값을 추출하는 과정에서 우리는 정책을 symmetric learning 방법으로 업데이트 한다. 이 때, 로봇의 관측 값 및 행동 값을 x축 방향으로 대칭시킨다. 그림 (4)를 보면 대칭 학습 시에는 로봇 조인트의 인덱스가 변환된 것을 확인할 수 있다. 로봇의 앞 왼쪽 다리의 조인트부터 인덱스를 순서대로 매기면 왼쪽 앞다리, 오른쪽 앞다리, 왼쪽 뒷다리, 오른쪽 뒷다리 순서로 부여된다. 이 인덱스 순서를 $[0, 1, 2, \cdots 9, 10, 11]$, 총 12개로 지정이 가능하고, 우리는 진행 방향에 대칭을 하기 때문에 이 조인트 인덱스가 오른쪽 앞다리, 왼쪽 앞다리, 오른쪽 뒷다리, 왼쪽 뒷다리 순서로 변환된다. $[3,4,5, \cdots 6,7,8]$ 순서로 지정이 된다. 또한 roll 방향으로 회전하는 hip 조인트의 경우, 왼쪽과 오른쪽이 대칭 변환되면 회전하는 방향도 +/- 로 반전이 되므로, 대칭성을 유지하기 위해 관측값 o_t에 -1의 배수 (multiplyer)를 곱하여 x축 방향

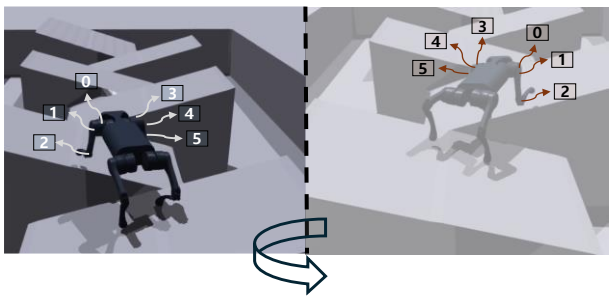


그림 4. 행동-상태 와 대칭 행동-상태.

Fig. 4. State-action set and symmetric state-action set.

표 1. 보상 함수.
Table 1. Reward function.

Reward	Equation	weight
Goal velocity	$\min\left(\langle v, \widehat{d_w}\rangle, v_{\text{cmd}}\right)$	1.5
Yaw velocity	$e^{-\lvert h_{target}-h\rvert}$	0.5
Z velocity	v_z^2	-1.0
Roll-pitch velocity	$\lvert\omega_{x,y}\rvert^2$	-0.05
Orientation	$\lvert R_{base}^T \cdot g\rvert^2$	-1.0
Joint position	$\lvert q_t^{nominal} - q_t\rvert^2$	-0.04
Joint acceleration	$\ddot{q}^2$	-2.5e-7
Action rate	$\lvert a_{t-1} - a_t\rvert$	-0.1
Energy	$\lvert\tau_t\rvert^2$	-1e-5
Hip pos	$\left(q_{hip}^{nominal} - q_{hip}\right)^2$	-0.5
Foot stumble	$\sum \mathbb{1}_{\lvert f_{foot,x,y}\rvert > c_{foot,x,y}}$	-1.0
Foot edge	$\sum \text{o1}\left(x_{\text{foot}}^{x,y} \lvert \text{mask}_{x,y}^{\text{edge}} \& C_{\text{contact}}^{\text{foot}}\right)$	-1.0
Collision	$\mathbb{1}_{collision}$	-10.0
Delta torque	$\lvert\tau_t - \tau_{t-1}\rvert^2$	-1e-7

으로 대칭된 관측값 $f(o_t)$로 변환한다. 이와 유사하게 주변 지형 정보의 높이 값 m_t에 대응되는 각 y 좌표 또한 x축 대칭 변환을 적용하여 새로운 대칭 지형 정보 $m_{t,s}$를 생성한다. 이렇게 매핑된 상태들은 그림 (2)의 대칭 상태(symmetric state)에 해당하고 대칭 액터-크리틱의 인풋으로 들어가게 된다. 대칭 네트워크는 대칭 변환되기 전의 액터-크리틱의 뉴럴 네트워크와 구조적으로 동일하나, 입력과 출력이 대칭 관계에 따라 적절히 변환된다. 이 대칭 네트워크는 손실함수를 식 (14)와 같이 변형하여 정책을 업데이트 한다.

다음으로 파쿠르 태스크에서 사용된 보상 함수에 대해 논의하고자 한다. 파쿠르 태스크의 보상함수는 기존의 연구 [7]를 기반으로 디자인되었고, 구체적인 구성 요소는 표 (1)에 정리되어 있다. 첫 번째 두 개 항은 reward task 항으로 로봇이 지향되는 방향의 보상함수이다. 여기서는 임의로 로봇 진행 방향에 목표지점(goal point)을 지정하여 로봇이 목표지점을 향하게 직진하여, 빠르게 파쿠르 태스크를 수행하도록 하였다. 첫 번째 goal velocity 함수는 로봇이 goal point 위치를 따라가도록 하는 항이고 두 번째 항인 Yaw velocity은 각 장애물과 로봇 사이의 방향 각도가 최소가 되도록, 즉 방향을 따라가도록 하는 보상 함수이다. 밑에 소개되는 다른 항들 경우에는 로봇이 과도하게 행동하는 것을 방지하는 규제 항으로서, 조금 더 부드러운 모션 및 안전한 모션을 만드는 데 기여한다. 특히, 파쿠르와 같은 행동은 장애물을 넘나드는 모션이 많은데, 지형을 제대로 짚지 못하면, 다리가 빠져버리는 위험한 상황이 나올 수 있다. 따라서 Foot edge 항을 통해, 로봇의 발이 장애물의 edge 근방에 위치해 있지 않고 어느 정도의 거리를 둔 안전범위 내에 존재하도록 한다. 그리고 collision의 보상 함수 스케일을 -10.0의 값으로 설정하여 장애물과의 충돌을 줄이려고 하였다. 또한 학습 시 지형을 커리큘럼으로 학습시켜서, 점차 낮은 레벨의 지형부터 어려운 지형을 학습하도록 하였다. 그림 (3)의 (a)에 해당하는 허들 지형은 [0.1, 0.4] m 로 높이 범위를 지정하여 단계별로 학습시켰고, (b)의 계단 지형은 [0.1, 0.45] m, (c)의 경우에는 최고 기울어진 높이를 [0, 0.25] m로 설정하였다. (d)의 갭이 있는 지형은 [0.1, 0.6] m의 갭의 범위로 두어 커리큘럼 학습을 진행하였다.

III. 시뮬레이션 실험

이번 장에서는 대칭 손실 함수를 고려한 PPO+MSL, PPO+PSL의 방법과 기존의 PPO방법에 대해서 사족보행 파쿠르 태스크에서의 성능을 비교하려고 한다. 크게 3가지 관점에서 실험을 비교할 것이다. 첫 번째는 학습의 수렴속도 및 모델의 최대 성능치, 두 번째는 외란에 대한 로봇의 강건함, 세 번째는 경험하지 못한 환경에서의 보행의 안정성이다. 본 실험은 시뮬레이션 Isaac Gym에서 NVIDIA RTX 4080을 사용하여 수행되었다. 4,096개의 환경에 대해서 총 8,000번의 iteration을 돌렸고 평균 6~7시간이 소요되었다.

1. 학습의 수렴속도

그림 (5)를 통해 iteration별 PPO, PPO + PSL, PPO + MSL의 평균 리턴 값을 비교할 수 있다. 초기 학습 단계에서 가장 빠르게 특정 수렴 구간에 도달한 방법은 PPO + PSL이고 학습 iteration이 500번 정도 되었을 때는 리턴 값이 20정도에 도달하였다. 다른 대칭 학습 방법인 PPO + MSL의 경우에는 17.5 근방, 기본 PPO는 16 근방의 평균 리턴 값을 가진다. MSL 방법론을 접목한 경우에는, 약 9.4%의 상승, PSL의 경우에는 약 25.0%정도가 상승하였다. 같은 반복 횟수의 training 동안에 대칭적 학습 방법이 조금 더 높은 보상값을 가지는 것을 확인할 수 있을 뿐만 아니라 빠른 속도로 수렴하는 것을 확인할 수 있다. 이는 샘플링 효율성 관점에서 매우 긍정적이다. 더 적은 샘플을 가지고 같은 성능 값 혹은 그 보다 나은 성능 값을 가질 수 있기 때문에, 학습 시간적 관점에서 경제적이다. 또한 최종 리턴 값도 대칭 학습을 적용한 방법이 높은 값을 보여주었기 때문에, 모델의 성능치를 최대화할 수 있었다

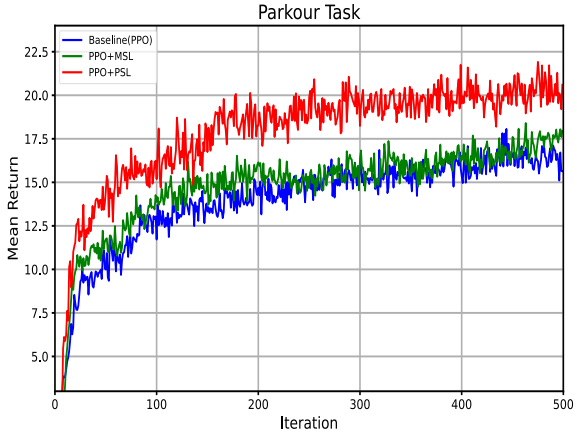


그림 5. 기본 PPO, PPO+MSL, PPO+PSL 간의 학습 곡선 비교.
Fig. 5. Comparison of training curves between PPO, PPO + MSL, PPO + PSL.

표 2. 학습 분포 내 환경(in-distribution, ID)과 분포 외(out of distribution, OOD)에서의 평균 성공률(SR) 및 평균 보상 비교.
Table 2. Comparison of mean Success Rate (SR) and mean reward on the in-distribution (ID) environment and out of distribution (OOD) environment.

Method	ID Mean SR (%)	OOD Mean SR (%)	ID Mean reward	OOD Mean reward
PPO	87.3	70.2	22.60	13.46
PPO+MSL	87.4	72.8	22.90	13.81
PPO+PSL	88.5	76.7	23.14	14.37

표 3. 충격량이 작용했을 때 Success Rate (SR) 및 평균 보상 지표.
Table 3. Success Rate (SR) and mean reward when impulse is applied to the robot.

Impulse (N · s)	Method	SR (%)	Mean Reward
50	PPO	58.5	17.48
	PPO + MSL	59.8	17.61
	PPO + PSL	63.1	18.22
80	PPO	52.7	15.56
	PPO + MSL	54.3	15.91
	PPO + PSL	57.1	16.51
120	PPO	45.9	13.22
	PPO + MSL	47.1	13.54
	PPO + PSL	49.3	14.11

2. 외란에 대한 강건성

로봇의 대칭성 학습에 의한 강건함을 보여주기 위해 외란이 작용했을 때, 파쿠르 테스크를 잘 수행하는 지 보여주려고 한다. 실험환경은 다음과 같다. 우선 로봇이 진행하는 방향에 수직인 y축 방향으로 50, 80, 120 N · s 의 충격량을 가한다. 이는 로봇 무게가 12.45kg임을 고려하면 보행안정성을 해치기에 충분한 수치이다. 우리는 여기서 성능 지표로 중간의 목표 지점과 최종 목표 지점까지 잘 도착했는 지를 판단하는 Succes Rate (SR)와 시나리오가 진행되고 있는 과정에서의 보상 함수의 평균을 선택했다. 그림 (3)의 4개의 훈련된 지형에 대해서 총 256개의 로봇을 병렬적으로 테스트하여 평균값을 구하였다. 표 (3)에서 보면, 50 N · s 의 충격량이 가해졌을 때 PPO 방법에 대조되어 PPO+ MSL은 성공률은 59.8%로 약 2.2%, 평균 보상값은 17.61로 0.74% 상승하였다. PPO+PSL 방법론은 성공률은 7.86% 평균 보상값은 4.23% 로 상승하였다. 또한 추가적으로 더 큰 충격량인 80 N · s 이 가해졌을 때 PPO+MSL의 성공률은 54.3%로 약 3.04%, 평균 보상값은 15.91로 2.25% 상승하였다. PPO+PSL 방법론의 성공률은 8.35% 평균 보상값은 6.10% 로 상승하였다. 마지막으로 120 N · s 의 충격량이 가해졌을 때 MSL의 방법론은 성공률은 약 2.61%, 평균 보상값은 2.42% 상승하였다. PPO+PSL 방법론은 성공률은 7.40% 평균 보상값은 6.73% 로 상승하였다. 충격량이 강해질수록 로봇의 전체적인 성공률과 평균 보상값은 줄어드는 것을 알 수 있다. 그리고 모든 충격량 실험의 경우에 기존 PPO보다 MSL이나 PSL을 결합한 대칭 정책이 더 높은 지표를 보이는 것을 알 수 있다. 이는 대칭기반의 정책이 외란에 대해서도 강건하다는 것을 보여준다. 로봇의 힘 분배가 균등하게 이루어지면서 안정성을 유지할 수 있었다.

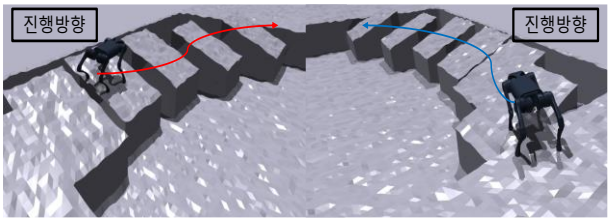


그림 6. 분포 외 (out of distribution, OOD) 실험 환경.
Fig. 6. Out of distribution (OOD) experiment environment.

3. 분포 외 환경에 대한 적응성

모델의 일반화 능력을 평가하기 위해 분포 외(out of distribution, OOD)환경에서의 성공률(SR, Succes Rate) 및 평균 보상을 측정하였다. OOD 환경은 학습 중에 경험하지 못한 상황을 나타내어 모델의 적응성에 대해서 실험해보았다. 그림 (6)은 분포 외 테스트 환경을 나타내는 그림이다. 두 가지 분포 외 지형에서 테스트를 진행하였고, 두 지형 모두 한 방향으로만 곡선이 형성되어있다. 뿐만 아니라 그림 (3)의 4개의 테스트 환경에서도 외란이 작용하지 않았을 때의 성공률과 평균 보상 값을 비교해보았다. 이 테스트 환경이 훈련 환경과 다른 점은 4개의 환경에 대한 난이도를 중간 단계로 고정하였다는 점이다. (a)의 허들 지형은 0.2m의 높이, (b)의 계단 지형은 0.25 m, (c)의 경우에는 최고 기울어진 높이를 0.1m로 두었다. (d)의 갭이 있는 지형은 0.3 m의 갭의 범위로 설정하였다. 표 (2)에서 분포 내 환경(in-distribution, ID)에서의 성공률과 보상 값을 비교하면 MSL 방법은 각각 약 0.11% 상승, 1.33%의 상승을 보여주었다. PSL 방법도 마찬가지로 기존의 PPO보다 약 1.37%, 2.39%로 증가하였다. 극명한 차이는 OOD에서 보이는데, PPO + MSL 방식은 성공률과 평균 보상에서 약 3.70%, 2.60%로 증가하였고, PPO + PSL은 약 9.26%, 6.76%로 증가하였다. 이를 통해 대칭 기반 정책이 로봇이 새로운 지형에 적응하는 능력을 향상시켜 경험하지 못한 다양한 상황에서 안정적인 성능을 발휘할 수 있음을 의미한다. 특히 PPO + PSL 방식이 기존과 대비하여 높은 성능 지표를 보여주었으며, 이는 MSL 방법처럼 단순히 대칭 액터와 액터의 행동 값에 대한 MSE를 줄이기보다는 식 (11)~(13)과 같이 이전 정책과 비교하면서 일정 범위로 클램핑하여 업데이트 하는 것이 적응성 및 안정성 측면에서 더 나은 결과를 보여주는 것을 확인하였다.

IV. 결론

본 연구에서는 대칭 손실을 적용한 PPO+MSL, PPO+PSL 방법이 기존의 PPO 방법에 비해 사족보행 로봇의 파쿠르 태스크 수행에서 어떠한 성능 향상을 가져오는지 실험적으로

확인하였다. 시뮬레이션 실험 결과를 통해 수렴 속도, 외부 충격량에 대한 강인함, 분포 외 환경에서의 성공률 및 보상 등 다양한 측면에서 성능을 평가하였다. 첫째, PPO+PSL 방식은 기존 PPO 대비 평균 보상값이 큰 값으로 수렴한다. MSL 방법론을 접목한 경우에는, 약 9.4%의 상승, PSL의 경우에는 약 25.0% 정도가 상승하였다. 또한 수렴 속도가 빠르고 학습 효율성이 높아, 더 적은 반복 횟수로 높은 성능을 달성할 수 있음을 확인하였다. 이는 샘플 효율성 면에서 큰 장점을 가진다. 둘째, 외부 충격량이 가해진 상황에서도 PPO+MSL 및 PPO+PSL 방식이 PPO보다 높은 성공률과 보상 값을 기록하며, 강인함 측면에서 개선된 성능을 보였다. 이는 대칭성을 활용한 학습 방법이 로봇의 균형 유지와 안정성을 강화하여, 임의의 방향에서의 외란이 작용하는 환경에서도 향상된 성능을 발휘할 수 있음을 시사한다. 셋째, OOD 환경에서도 대칭 손실을 적용한 PPO+PSL 방식은 기존 PPO에 비해 높은 성공률과 보상을 기록하며, 뛰어난 일반화 성능을 보였다. 특히, PPO+PSL 방식은 분포 외 환경에서의 평균 성공률이 9.26% 상승하였고, 평균 보상 또한 6.76% 증가하여 예측하지 못한 환경에서도 높은 성능을 유지하였다.

결론적으로, 본 연구의 결과는 대칭 손실을 활용한 강화학습 방법론이 사족보행 로봇의 강건성 및 적응성을 향상시키는 데 효과적인 접근임을 보여주었다. 특히 PSL 방법론을 접목한 경우에 두 가지 방법보다 나은 결과를 보여주었다. PPO+PSL은 이전 정책의 행동 확률 분포를 기준으로 현재 정책의 확률 분포 변화를 신뢰 구간 내에서 조절함으로써 학습의 효과를 입증하였다. 이러한 접근은 대칭 상태에서의 학습을 점진적으로 진행시키고, 급격한 변화로 인한 불안정을 방지하여 정책의 안정성을 높이는 역할을 한다.

향후 이 학습 방법론을 실제 로봇에 적용하여, 검증하는 것을 최종 목표로 하고 있다.

REFERENCES

[1] A. Kumar, Z. Fu, D. Pathak, and J. Malik, “Rma: Rapid motor adaptation for legged robots,” arXiv preprint arXiv:2107.04034, 2021.
doi: https://doi.org/10.48550/arXiv.2107.04034

[2] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, “Rapid locomotion via reinforcement learning,” *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572-587, 2024.
doi: https://doi.org/10.48550/arXiv.2205.02824

[3] Z. Fu, X. Cheng, and D. Pathak, “Deep whole-body control: learning a unified policy for manipulation and locomotion,” *Proceedings of the 6th Conference on Robot Learning,* PMLR, pp. 138-149, 2023.
doi: https://doi.org/10.48550/arXiv.2210.10044

[4] S. Jeon, N.-H. Kwon, S.-M. Ha, and J.-Y. Kim, “Integrated walking control strategy with model predictive control and whole–body control using deep learning–based state estimator,” *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 10, pp. 1139-1146, 2024.

[5] B. Kim, W. Yoon, G. Kwon, M. Park, and N. K. Kwon, “Stair detection algorithm using depth line profile for the autonomous inter-floor movement of the multi-legged robot,” *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 8, pp. 780-786, 2024.

[6] Z. Zhuang, S. Yao, and H. Zhao, “Humanoid parkour learning,” arXiv preprint arXiv:2406.10759, 2024.
doi: https://doi.org/10.48550/arXiv.2406.10759

[7] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, “Extreme parkour with legged robots,” *2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, pp. 11443-11450, 2024.
doi: https://doi.org/10.48550/arXiv.2309.14341

[8] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, “Robot parkour learning,” *Conference on Robot Learning (CoRL)*, 2023.
doi: https://doi.org/10.48550/arXiv.2309.05665

[9] W. Yu, G. Turk, and C. Karen Liu, “Learning symmetric and low-energy locomotion,” *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1-12, 2018.
doi: https://doi.org/10.48550/arXiv.2309.02711

[10] M. Abreu, L. Paulo Reis, and N. Lau, “Addressing imperfect symmetry: a novel symmetry-learning actor-critic extension,” *Neurocomputing*, vol. 614, 128771, 2025.
doi: https://doi.org/10.48550/arXiv.2309.02711

[11] M. Kasaei, M. Abreu, N. Lau, A. Pereira, and L. P. Reis, “A CPG-based agile and versatile locomotion framework using proximal symmetry loss,” arXiv preprint arXiv:2103.00928, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00928

[12] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, “Learning to walk in minutes using massively parallel deep reinforcement learning,” *Conference on Robot Learning*, PMLR, pp. 91-100 2022.
doi: http://dx.doi.org/10.1007/s12555-020-0982-8

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv preprint arXiv:1707.06347, 2017.
doi: https://doi.org/10.48550/arXiv.1707.06347

[14] G. Ji, J. Mun, H. Kim, and J. Hwangbo, “Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4630-4637, April 2022.
doi: https://doi.org/10.1109/LRA.2022.3151396

김 동 주

2024년 부산대학교 기계공학부(공학사). 2024년~현재 부산대학교 전기전자공학과 석사과정 재학 중. 관심분야는 강화학습, 보행제어.

박 지 훈

2021년 부산대학교 전자공학과(공학사). 2023년 한국과학기술원 전기전자공학부(공학석사). 2024년~현재 부산대학교 전기전자공학과 박사과정 재학 중. 관심분야는 강화 학습, 보행제어.

이 인 호

2009년 한국과학기술원 기계공학과(공학사). 2011년 한국과학기술원 기계공학과(공학석사). 2016년 한국과학기술원 기계공학과(공학박사). 2020년~현재 부산대학교 전기자공학과 교수. 관심분야는 로보틱스, 기계시스템 자동화, 보행로봇.