

사족보행 로봇의 선박 내 활용을 위한 보행 제어 학습

Learning Gait Control for Quadrupedal Robots Operating on Ships

구 본 석^{1b}, 정 종 대^{1*}

(Bonseok Gu¹ and Jongdae Jung^{1,*})

¹Department of Autonomous Vehicle System Engineering, Chungnam National University

Abstract: To deploy quadruped robots for inspection tasks on ships, the locomotion controller must maintain stability under the six degrees-of-freedom motion of the ship. In this work, we propose a robust locomotion control policy trained using the proximal policy optimization algorithm to maintain robustness against dynamic disturbances from ship motion. To achieve this, a ship motion simulation platform is developed in Isaac Sim, where periodic roll and pitch motions are applied to simulate realistic ship dynamics. In addition, a reward function to improve gait stability and velocity tracking under dynamic disturbances is developed, and the training hyperparameters are tuned to enhance robustness. The proposed controller is compared with controllers trained on flat and unstructured rough terrain. The simulation results show that all three policies stably converged, achieving stable walking on the terrains used for training each controller. However, under disturbance conditions such as ship motion, the proposed policy maintained a stable posture and reliably tracked the commanded speed, whereas the baseline controllers failed to maintain a stable performance. These results demonstrate the effectiveness of the proposed controller for onboard inspection tasks.

Keywords: Reinforcement learning, quadruped robot, dynamic ground, locomotion control

1. 서론

최근 건조되는 선박 내에는 복잡하고 다양한 장비들이 설치되어 있으며, 이들의 신속하고 정확한 점검은 안전 운항과 유지보수 측면에서 필수적이다[1]. 그러나 선박 내부는 좁고 파도에 의한 선박의 운동 등 불안정한 환경적 특성이 존재하여, 기존의 사람이 직접 점검하는 방식은 접근성, 안전성과 효율성에 한계가 있다. 이에 따라, 사족보행 로봇을 활용한 자동화된 장비 점검 기술이 주목받고 있다[2-3].

사족보행 로봇은 높은 기동성과 다양한 지형 적응력이 강점이지만, 기존의 사족보행 로봇의 제어에 관한 연구는 주로 사족보행 로봇을 육상에서 활용하는 상황에 대해서 이루어지고 있다. 따라서 기존 연구들은 주로 정적인 지형에서 사족보행 로봇의 보행 제어 성능을 개선해왔다[4-6]. 그러나 선박 내부와 같이 움직이는 지면에서 사족보행 로봇을 운용하기 위해서는 선박의 움직임과 같은 동적인 지면의 운동에 대한 고려가 필수적이다.

사족보행 로봇의 보행 제어 기법은 크게 모델 기반 제어 기법과 학습 기반 제어 기법으로 나눌 수 있다. 사족보행 로봇은 주로 고전적인 모델 기반 제어 기법이 사용되어 왔으나, 최근 시뮬레이션 고도화와 강화학습 알고리즘의 발전으로 사족보행 로봇의 보행 제어에 강화학습을 성공적으로 적용한 연구가 증가하고 있다[7-9]. 모델 기반 제어 기법과 달리, 이러한 접근은 모델링에 필요한 시간과 노력을 줄여 주며,

특히 숙련되지 않은 사용자에게 까다로운 운동학적 자세 제어를 더욱 수월하게 한다[10]. 그리고 학습 기반 제어 기법의 경우 모델 기반 제어 기법에 비해 일반적으로 외란 저항 능력이 크고 에너지 소비가 적음을 검증한 연구 결과가 존재한다[11-12].

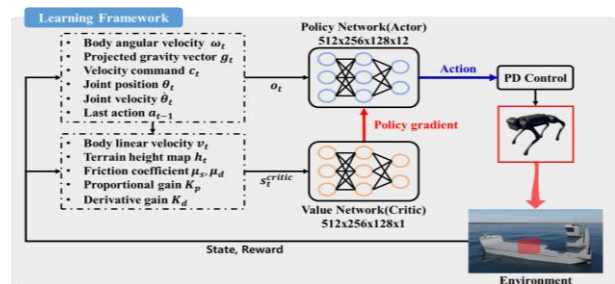


그림 1. 요동하는 선박 환경에서 안정적인 보행을 위한 강화 학습 기반 제어기의 학습 및 검증 프레임워크.

Fig. 1. Reinforcement learning-based training and validation framework for stable locomotion control in oscillating ship environments

* Corresponding Author

Manuscript received November 15, 2025; revised December 2, 2025; accepted January 5, 2026

구본석: 충남대학교 자율운항시스템공학과 대학원생(bonseokgu@o.cnu.ac.kr, ORCID[®] 0009-0007-7651-578X)

정종대: 충남대학교 자율운항시스템공학과 조교수(jdjung@cnu.ac.kr, ORCID[®] 0000-0001-6591-213X)

※ 이 논문은 2025년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (2410010873, 2025년 산업혁신인재성장 지원사업).

※ 본 연구는 충남대학교 학술연구비에 의해 지원되었음.

본 연구에서는 선박 내 지면의 움직임에도 안정적으로 작동할 수 있도록 학습 기반 보행 제어기를 설계하고, 이를 시뮬레이션 환경 내에서 검증하였다. 특히 요동하는 선박 위에서 사족보행 로봇이 안정적으로 속도 명령 추종 및 경유점 추종 성능에 대한 시뮬레이션 검증을 진행하였다.

II. 관련 연구

1. 보행 로봇의 동적 지면 적응 제어

동적 지면에 대한 보행 로봇의 제어에 관한 연구는 많지 않으나 최근 일부 연구에서 보행 로봇의 자세 안정성을 확보하기 위한 학습 기반 제어 기법들을 제안하고 있다. 이동 플랫폼 위에서 발생하는 병진 및 회전 운동에 대응하기 위해 로봇의 자체 센서로부터 얻은 관절의 위치와 속도, IMU (Inertial Measurement Unit) 센서의 가속도 및 각속도 그리고 이전 시점의 행동 정보를 통해 로봇의 선속도, 질량, 마찰, 이동 플랫폼의 선속도 및 각속도를 추정하는 이중 상태 추정기를 개발하고, 이를 활용한 동적 지면에서의 자세 제어기 학습 프레임워크를 제안한 연구가 존재한다[13]. 또한 이동 플랫폼에 IMU를 부착하여 지면의 움직임을 측정하고, 이를 정책 신경망의 입력으로 반영하는 방법을 통해 자세 보정을 수행한 학습 기반 제어 프레임워크를 개발한 연구 사례가 있다[14]. 하지만 이러한 선행 연구들이 로봇의 정적 안정성에 초점을 맞춘 반면에, 실제 선박에서 작업하게 될 보행로봇은 정적 안정성뿐만 아니라 동적 안정성도 요구된다. 본 연구에서는 요동하는 선박과 같은 동적 지면 환경에서 자세 안정화와 함께 안정적인 이동도 가능한 보행 제어기를 설계하였다.

III. 보행 제어기 학습 프레임워크

1. 강화학습 알고리즘

1.1 강화학습 개요

강화학습은 에이전트가 환경과의 상호작용을 통해 누적 보상을 최대화하는 최적 정책을 학습하는 방법이다. 이때 에이전트는 매 시점 t 에서 환경의 상태 s_t 를 관측하고, 정책에 따라 행동 a_t 를 결정한다. 환경은 선택된 행동에 대한 결과로 다음 상태 s_{t+1} 를 반환하며, 에이전트는 이를 이용해 누적 보상의 기대값을 최대화하도록 정책을 갱신한다. 강화학습 알고리즘은 일반적으로 에이전트와 환경이 상호작용하는 과정을 MDP (Markov Decision Process)로 표현하여 수학적으로 정의한다. MDP는 $M = (S, A, P, R, \gamma)$ 로 구성된다. S 는 에이전트의 상태 집합, A 는 에이전트가 정책을 통해 취하는 행동의 집합, P 는 에이전트가 특정 상태 s 에서 행동 a 를 선택했을 때 다음 상태 s' 로 전이될 확률의 집합이다. R 은 에이전트가 상태 s 에서 행동 a 를 취했을 때 받는 보상의 집합이며, γ 는 보상을 누적하여 미래에 얻을 수 있는 보상의 비중을 결정하는 할인율이다.

1.2 PPO 알고리즘

본 연구에서는 연속적인 제어 입력을 안정적으로 학습하기 위해 PPO (Proximal Policy Optimization) [15] 알고리즘을 적용하였다. PPO 알고리즘은 정책 기반 강화학습 기법 중 하나로, 기존의 TRPO (Trust Region Policy Optimization) [16] 알고리즘에서 계산 복잡도를 줄이면서도 안정적인 정책 업데이트

를 가능하게 한 방법이다.

정책 기반 방법에서는 에이전트의 행동 정책 $\pi_\theta(a|s)$ 을 매개변수 θ 로 표현하고, 환경으로부터 얻은 보상 R_t 를 통해 기대 누적 보상을 최대화하도록 학습한다. PPO 알고리즘은 이 과정에서 정책이 급격하게 변경되어 학습이 불안정해지는 문제를 방지하기 위해, 식 (1)과 같은 클리핑 기법을 도입하였다. $r_t(\theta)$ 는 새로운 정책과 기존 정책의 변화율이며 식 (2)와 같이 계산된다. $\hat{A}_t$ 는 이득 함수 추정값, ϵ 은 클리핑 파라미터로 허용 가능한 정책 변경 범위를 나타낸다. $\hat{E}_t$ 는 시간 t 에서의 기대값 연산을 의미하며, 학습 과정에서 미니배치 단위의 샘플 데이터를 수집하고, 이에 대한 평균을 계산한다. $\text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)$ 는 정책 변화율을 클리핑 파라미터로 제한하여 일정 비율보다 커지거나 작아지는 것을 방지하는 역할을 한다. 이 식을 통해 PPO 알고리즘은 정책이 과도하게 업데이트되는 것을 제한하여, 안정적인 학습을 유지하면서도 수렴 속도를 향상시킨다.

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

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \quad (2)$$

또한 PPO 알고리즘은 학습 과정에서 가치 함수 손실과 엔트로피항을 함께 최적화하여 탐색과 수렴의 균형을 유지한다. 전체 목적함수는 식 (3)과 같이 정의된다. 여기서 L^{VF} 는 가치함수 손실, $S[\pi_\theta]$ 는 정책의 엔트로피이고, c_1 은 가치함수 손실 가중치이며 가치함수의 학습 강도를 결정한다. 가중치가 커질수록 정책 개선 속도가 저하될 수 있으며 이로 인해 수렴이 느려지거나 성능 향상이 제한될 수 있다. 또한 작아질수록 가치함수 추정의 분산이 증가하며 정책 개선이 불안정해질 수 있다. c_2 는 엔트로피 가중치 계수이며 탐색 강도를 조절한다. 엔트로피 가중치가 커질수록 불필요한 탐색으로 정책의 변동성이 증가하고 수렴이 지연될 수 있다. 반대로 작아질수록 탐색이 부족해 초기에 수렴하여 지역해에 빠질 가능성이 증가해 성능이 감소할 수 있다. 가치 함수 손실은 식 (4)와 같이 정의되며, 가치 함수의 예측값 $V_\theta(s_t)$ 와 현재 상태에서 얻을 것으로 추정되는 누적 보상의 추정값 V^{target} 의 MSE (Mean Squared Error)로 정의된다.

$$L^{\text{PPO}}(\theta) = \hat{E}_t[L^{\text{CLIP}} - c_1 L^{\text{VF}}(\theta) + c_2 S[\pi_\theta](s_t)] \quad (3)$$

$$L^{\text{VF}} = (V_\theta(s_t) - V^{\text{target}})^2 \quad (4)$$

2. 시뮬레이션 환경 구성

2.1 시뮬레이터 선정

강화학습 기반 제어기 학습에서는 방대한 양의 데이터를 반복적으로 수집하고 이를 통해 정책을 개선해야 한다. 그러나 실제 로봇 플랫폼에서 이러한 데이터를 취득하는 과정은 시간과 비용이 많이 소요될 뿐 아니라, 하드웨어 마모 및 충돌 위험이 존재한다. 이에 본 연구에서는 제어기 학습 효율을 높이기 위해 물리 기반 시뮬레이터를 활용하였다.

시뮬레이터는 가상 환경에서 로봇의 동역학, 센서 모델, 접촉 등을 정밀하게 재현함으로써 실제 환경에서 발생할 수 있는 다양한 상황을 재현할 수 있다. 강화학습의 일반화 성능을 확보하기 위해서는 이러한 시뮬레이션 환경이 실제 물리

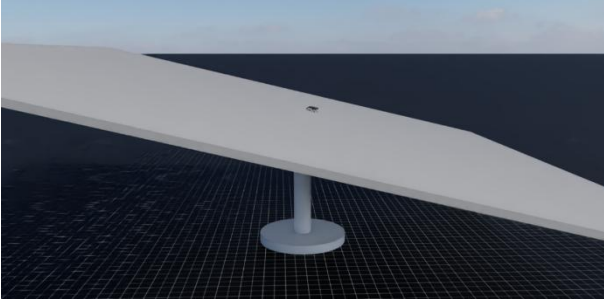


그림 2. 선박 운동 모사 플랫폼.

Fig. 2. Ship motion simulation platform.

특성과 유사하게 구성하는 것이 중요하다[17].

대인 로봇 시뮬레이터로는 Gazebo [18], PyBullet [19], Isaac Sim [20] 등이 있으며, 본 연구에서는 그중 NVIDIA Isaac Sim 을 채택하였다. Isaac Sim은 NVIDIA Omniverse 플랫폼 기반의 시뮬레이터로, GPU 기반 물리 엔진인 PhysX와 병렬 시뮬레이션 기능을 지원하여 대규모 강화학습 환경을 효율적으로 구성할 수 있다. 따라서 본 연구에서는 Isaac Sim을 활용하여 강화학습 기반보행 제어기 학습용 시뮬레이션 환경을 구성하였으며, 이를 통해 대량의 데이터를 효율적으로 수집하고 보행 제어기를 개발하였다.

2.2 선박 운동 모사 플랫폼 구성

선박의 경우 파도, 바람, 조류 등의 해양 환경에 의해 6자유도 운동을 한다. 사족보행 로봇을 선내 검사 로봇으로 이용하는 대형선의 경우, 배수량이 크고 복원력이 높아 일반적으로 일반적인 해양 조건에서 횡동요 최대 경사각은 $5^\circ \sim 15^\circ$ 정도로 유지되고 주기는 약 15~30초이다[21]. 따라서 이를 고려하여 그림 2와 같이 선박의 움직임을 모사하여 종방향과 횡방향으로 움직이는 플랫폼을 구성하고, 횡방향과 종방향으로 진폭 15° , 주기 15초로 운동하도록 설정하여 학습 지형으로 활용하였다.

3. 강화학습 환경 구성

3.1 관측 상태 구성

본 연구에서는 강화학습 제어기의 정책 학습을 위해 PPO 알고리즘의 액터-크리틱 구조를 기반으로 한 학습 환경을 구성하였다. 정책과 가치 함수는 동일한 환경 내에서 동시에 학습되며, 에이전트가 주어진 상태를 효과적으로 인식하고 평가할 수 있도록 관측 상태를 정의하였다. 또한 본 연구에서는 비대칭 액터-크리틱 구조를 채택하였다.

이는 정책과 가치 함수가 서로 다른 관측 정보를 입력 받도록 설계하는 방식으로, 크리틱 네트워크가 액터 네트워크보다 더욱 많은 환경 정보를 활용하여 상태의 가치를 보다 정확하게 평가할 수 있도록 하는 구조이다. 이와 같은 비대칭 구조는 학습 안정성을 향상시키고 정책이 더 일반화된 행동을 학습하도록 유도한다.

관측 상태 $\mathbf{o}_t$ 는 실제 사족보행 로봇에 탑재된 센서를 고려하여 구성하였다. 우선 사족보행 로봇에 탑재된 IMU로부터 추정할 수 있는 몸체의 각속도 $\boldsymbol{\omega}_t$ 와 중력 방향 벡터 $\mathbf{g}_t$, 속도 명령 $\mathbf{c}_t$ 와 각 관절의 모터 엔코더로부터 측정할 수 있는 관절의 위치 $\boldsymbol{\theta}_t$ 와 속도 $\dot{\boldsymbol{\theta}}_t$, 그리고 이전 시점의 액션 $\mathbf{a}_{t-1}$ 을 포함하도록 구성하였다.

표 1. 보상함수.

Table 1. Reward function.

Term	Equation (r_t)	Weight (ω_i)
Lin. velocity tracking	$\exp(-\frac{ \mathbf{c}_{xy} - \mathbf{v}_{xy} }{0.25})^2$	1.5
Ang. velocity tracking	$\exp(-\frac{ \mathbf{c}_z - \boldsymbol{\omega}_z }{0.25})^2$	0.75
Feet air time	$t_{air} - 0.5$	0.25
Linear velocity (z)	$\mathbf{v}_z^2$	-2.0
Angular velocity (xy)	$ \boldsymbol{\omega}_{xy} ^2$	-0.05
Joint torque	$ \boldsymbol{\tau} ^2$	-2.0e-4
Joint acceleration	$ \ddot{\mathbf{q}} ^2$	-2.5e-7
Action rate	$ \mathbf{a}_t - \mathbf{a}_{t-1} ^2$	-0.01
Orientation	$ \mathbf{g}_{xy} ^2$	-2.5
Body height	h^2	-30.0
Foot torque balance	$std(\boldsymbol{\tau}_{foot} ^2)$	-1e-4

크리틱 네트워크의 입력값인 $\mathbf{s}_t^{critic}$ 은 관측 상태 $\mathbf{o}_t$ 를 포함하고, 크리틱 네트워크에만 전달하는 특권 정보인 몸체의 속도 $\mathbf{v}_t$, 주변 지형 높이 정보 $\mathbf{h}_t$, 지면의 마찰 계수 μ_s, μ_k , 제어기의 PD (Proportional Derivative) 이득 K_p, K_d 를 추가적으로 제공하여 학습 안정성을 높였다.

이와 같은 비대칭 액터-크리틱 구조를 통해, 제어기는 시뮬레이션 환경에서 로봇의 동적 상태와 외부 환경 변화를 보다 정밀하게 인식하고, 강화학습 과정에서 안정적이고 효율적인 정책 학습이 가능하도록 하였다.

$$\mathbf{o}_t = [\boldsymbol{\omega}_t \quad \mathbf{g}_t \quad \mathbf{c}_t \quad \boldsymbol{\theta}_t \quad \dot{\boldsymbol{\theta}}_t \quad \mathbf{a}_{t-1}]^T \quad (5)$$

$$\mathbf{s}_t^{critic} = [\mathbf{o}_t \quad \mathbf{v}_t \quad \mathbf{h}_t \quad \mu_s \quad \mu_d \quad K_p \quad K_d]^T \quad (6)$$

또한 본 연구에서는 강화학습 정책이 각 관절의 목표 위치를 출력하고 PD 제어기가 관절의 목표 위치와 실제 관절의 위치와 속도 오차를 이용해 토크를 산출하는 구조를 사용하였다. PD 제어기는 구현이 단순하고 고주파 제어에서 안정적으로 동작하며, 접촉이 반복되는 보행 과정에서 임피던스 특성으로 충격과 외란을 완화해 매끄러운 토크 출력과 자세 안정성 확보에 유리하다[22]. 실제 사족보행 로봇의 관절 제어시 PD 제어기는 식 (7)과 같이 정의되며, $\mathbf{q}$ 는 관절의 위치, $\dot{\mathbf{q}}$ 는 관절의 속도이다. 액션 $\mathbf{a}$ 는 빠른 수렴을 위해 학습 초기에 로봇이 서있는 자세를 유지하며 학습을 진행하도록 하였다. 따라서 로봇이 지면에서 서있는 자세에서의 관절 위치인 $\boldsymbol{\theta}_{stand}$ 로부터의 상대 관절 위치를 추론하도록 학습되어 PD 제어기에 최종 관절 위치 $\boldsymbol{\theta}_{des}$ 를 명령으로 전달하도록 구성하였다.

$$\boldsymbol{\tau} = K_p(\boldsymbol{\theta}_{des} - \mathbf{q}) - K_d\dot{\mathbf{q}} \quad (7)$$

$$\boldsymbol{\theta}_{des} = \boldsymbol{\theta}_{stand} + \mathbf{a} \quad (8)$$

3.2 보상함수 구성

선박 모사 플랫폼에서 안정적으로 작동하는 사족보행 로

표 2. 학습 중 적용된 도메인 랜덤화 목록.

Table 2. Domain randomization parameters used during training.

Term	Range
Static Friction Coefficient	[0.3, 1.25]
Dynamic Friction Coefficient	[0.3, 1.25]
Add Mass (kg)	[-1.0, 3.0]
Initial Joint Position	[-1.0, 1.0]
Push Robot (m/s)	[-0.5, 0.5]

표 3. 학습 파라미터 구성.

Table 3. Training hyperparameter configuration.

Term	Value
ϵ	0.2
c_1	1
c_2	0.01
Learning Rate	1.0e-3
λ	0.99

봇의 보행 제어를 개발하기 위해 표 1과 같이 보상함수를 구성하였다. 보상함수는 양의 보상을 주는 부분과 음의 보상을 주는 부분으로 나누어 구성하였다. 속도 명령을 안정적으로 추종할수록 양의 보상을 주도록 구성하였으며, 보행 중 로봇의 발 끝이 지면에 닿지 않는 시간 t_{air} 을 측정하여 보행 패턴의 주기를 0.5초에 가까워지도록 구성하였다.

로봇이 안정적으로 보행하도록 몸체의 회전, 각 관절이 받는 힘 그리고 관절의 가속도를 최소화하도록 음의 보상을 가하였다. 또한 선박 모사 플랫폼에서 학습을 진행할 시 로봇은 안정적인 이동을 위해 지면에 가깝게 높이를 낮춰서 이동한다. 그러나 움직이는 지면의 중심축으로부터 멀어질수록 지면의 회전 운동에 의해 z 방향으로 큰 움직임을 보인다. 이러한 경우 로봇의 몸체가 지면과 부딪히는 횟수가 늘어나 안정적인 이동이 불가능하다. 따라서 로봇의 몸체가 지면으로부터 일정 높이를 유지하도록 보상을 가하고, 각 발에 가해지는 힘이 균형을 이루도록 보상함수를 구성하였다. 각 보상은 식 (9)와 같이 가중치 w_i 와 곱하여 각 보상의 비율을 설정하였다.

$$r_t(s_t, a_t) = \sum_{i=1}^n r_i w_i \quad (9)$$

3.3 도메인 랜덤화

강화학습 기반 제어를 시뮬레이션을 통해 개발하고 실제 환경에 적용하기 위해, 실제 환경의 불규칙적인 환경과 다양한 외란에 대한 고려가 필요하다[23]. 이를 고려하기 위해 도메인 랜덤화 기법을 적용하여 사족보행 로봇이 다양한 조건에서도 강건하게 작동하도록 바닥 면의 마찰, 로봇의 무게 및 초기 관절 위치를 일정 범위 내에서 변경하고 무작위 외란을 적용하여 학습을 진행하였다. 그에 대한 자세한 수치는 표 2에 나타내었다.

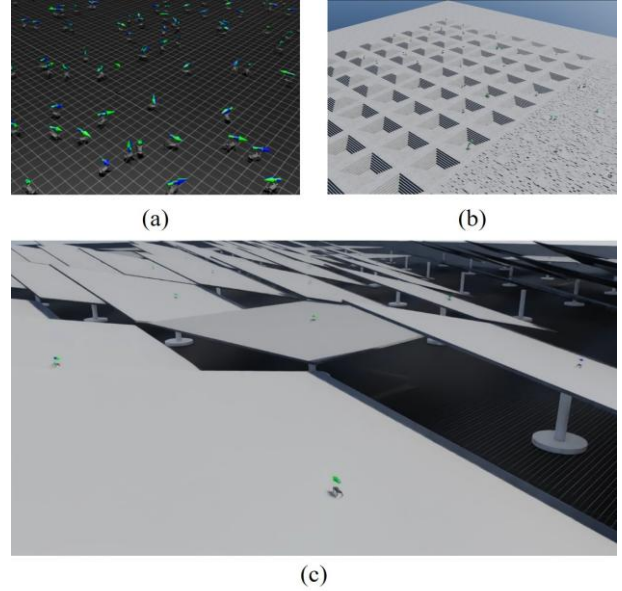


그림 3. (a) original and SAC, (b) baseline and (c) proposed 제어기 학습 시뮬레이션 환경.

Fig. 3. Simulation environment for training (a) original and SAC, (b) baseline, and (c) proposed controllers.

3.4 학습 파라미터

액터-크리틱 네트워크의 구조는 512x256x128의 은닉층으로 구성되어 있고 은닉층 사이의 활성화 함수는 ELU (Exponential Linear Units)로 설정했다. 학습에서 사용한 각 파라미터는 표 3에 나타내었다. ϵ 는 PPO 알고리즘의 과도한 정책 변화 방지를 위한 파라미터이고, λ 는 GAE (Generalized Advantage Estimation)에서 미래 시점에서 얻는 이득의 기여도를 조절하여 편향과 분산의 균형을 맞추기 위한 시간 감쇠 계수이다. 학습은 4096개의 에이전트를 병렬 학습 기법으로 동시에 학습하였다.

IV. 시뮬레이션

제어의 성능을 검증하기 위해 시뮬레이션 환경에서 기존 제어기들과 비교 실험을 수행하였다. 본 논문에서 비교 대상으로 사용된 제어기는 다음과 같이 정의한다.

- Original: 평탄한 지형에서 학습한 제어기
- Baseline: 계단 및 비정형 환경에서 학습한 제어기
- Proposed: 선박 운동 모사 플랫폼에서 학습한 제어기

또한 본 연구에서는 Off-Policy 알고리즘인 SAC (Soft Actor-Critic) [24] 알고리즘으로 학습한 제어기도 함께 비교하였다. 그림 3은 original, baseline, proposed, SAC 제어기의 훈련을 위한 시뮬레이션 환경이다. 그림 4는 original, baseline, proposed, SAC 제어기 훈련 중 속도 명령 추종 보상 값을 나타낸다. 그림 4에서 보듯이, PPO 알고리즘을 활용한 제어기 모두 약 2,000회 반복 후 안정적으로 수렴하여 제어기마다 학습에 활용한 지형에서 안정적인 보행 성능을 보이는 것을 확인할 수 있다. 반면 SAC 알고리즘을 활용한 제어기의 경우 z 방향 각속도 명령 추종 보상은 PPO와 유사한 수준까지 증가하였으나 xy 방향 선속도 명령 추종 보상은 학습 과정 내내 낮은

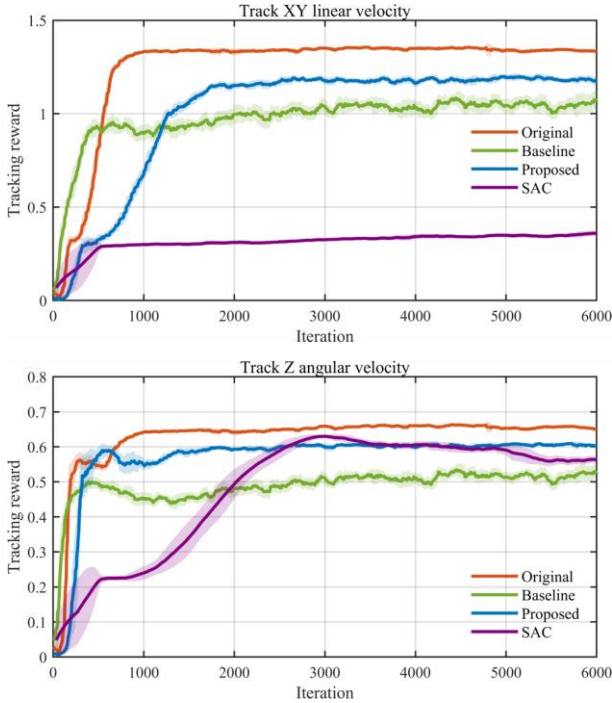


그림 4. Original, baseline, proposed, SAC 제어기의 학습 결과, xy축 선속도 추종 보상, z축 각속도 추종 보상.

Fig. 4. Training results of the original, baseline, and proposed controllers: xy linear velocity tracking reward, z-axis angular velocity tracking reward.

값을 유지하며 개선되지 못하는 경향을 보였다. 이는 선박 내 보행 제어 문제와 같이 외란이 크고 정밀한 보행 제어를 요구하는 상황에서 엔트로피를 최대화하는 방향으로 학습하는 SAC 알고리즘보다 PPO 알고리즘이 더 안정적으로 정책을 개선한다는 점을 나타낸다. 속도 추종 보상 값의 차이는 각 제어기의 학습 환경에 기인한 것으로 제어 성능 비교 평가에는 영향을 미치지 않는다. 제어기의 학습 시간은 3시간이며, Nvidia RTX 4090 단일 GPU를 사용하여 학습했다.

1. 속도 명령 추종 제어 성능 평가

제안된 제어기의 성능을 검증하기 위해, 그림 5-(a)와 같이 종방향으로 진폭 12°, 주기 15초의 운동을 하는 선박 환경을 구성하였다. 해당 환경에서 자세 유지 명령, 전진 속도 1.0 m/s 명령, 그리고 측면 속도 1.0 m/s 명령을 각각 적용하여 시뮬레이션 실험을 수행하였다. 또한 original, baseline 제어기와 비교하여 성능을 평가하였다.

그림 6은 선박 플랫폼에서 자세 유지 명령 하에 original, baseline, proposed 제어기의 xy 방향 신체 속도와 명령 속도를 비교한 그래프이며 파란색 화살표는 로봇의 몸체 속도를 나타낸다. original 제어기는 지면 움직임에 적응하지 못해 부딪치며 넘어진다. 반면 baseline, proposed 제어기는 속도 명령을 성공적으로 추종하며 원하는 자세를 유지한다.

그림 7은 1m/s 전진 속도 명령 하에 original, baseline, proposed 제어기의 xy 방향 몸체 속도와 명령 속도를 비교한 그래프이다. Original 제어기는 전진 속도 명령을 추종하지 못하고 선박의 움직임에 따라 속도가 변한다. 반면 baseline, proposed 제어기는 속도 명령을 안정적으로 추종함을 확인할



(a)



(b)

그림 5. 제안된 제어기의 성능 검증을 위한 시뮬레이션 환경. (a) 속도 명령 추종 성능 평가 환경. (b) 경유점 추종 성능 평가 환경.

Fig. 5. Simulation environments for evaluating the performance of the proposed controller, (a) velocity command tracking evaluation environment, (b) waypoint tracking evaluation environment.

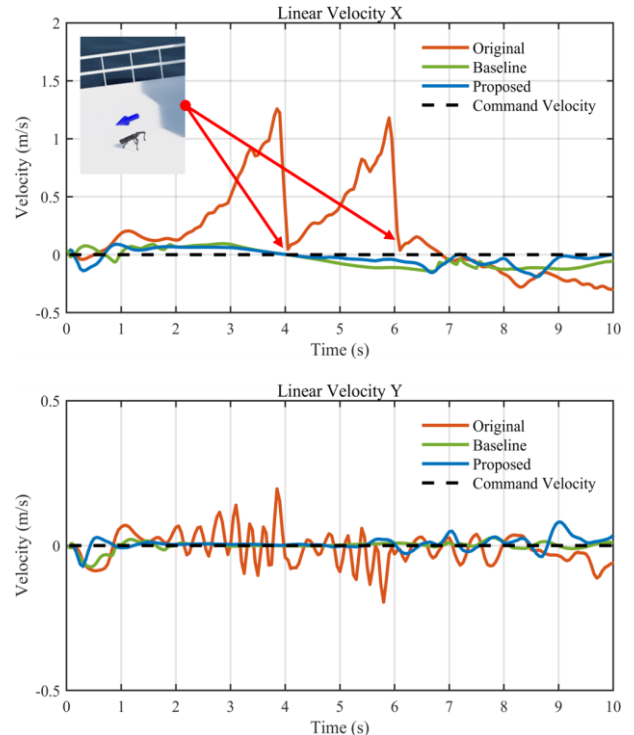


그림 6. 요동하는 선박 환경에서 original, baseline, proposed 제어기의 자세 유지 명령 성능 비교.

Fig. 6. Performance comparison of the original, baseline, and proposed controllers under posture maintenance in a ship motion environment.

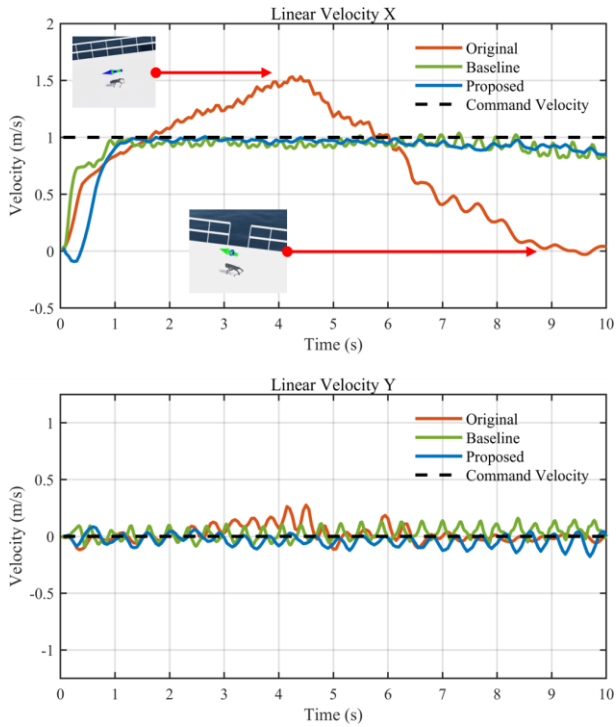


그림 7. 요동하는 선박 환경에서 original, baseline, proposed 제어기의 전진 속도 1 m/s 명령 성능 비교.

Fig. 7. Performance comparison of the original, baseline, and proposed controllers under a forward velocity of 1 m/s in a ship motion environment.

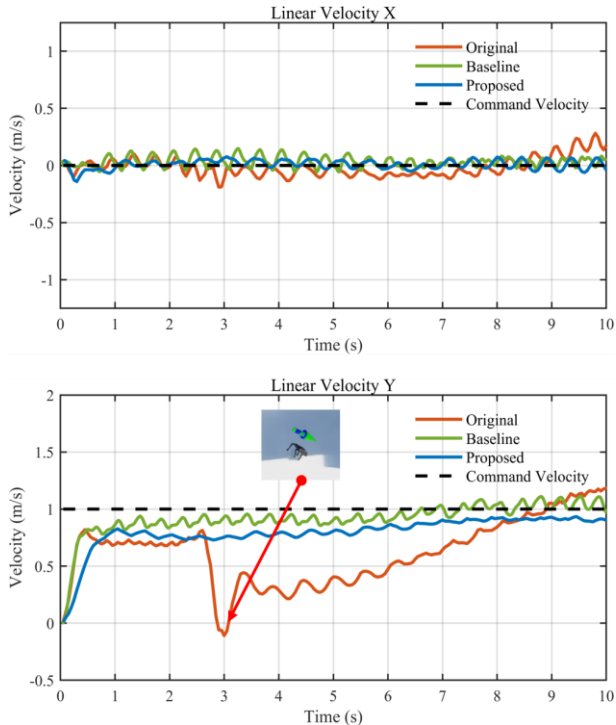


그림 8. 요동하는 선박 환경에서의 original, baseline, proposed 제어기의 횡방향 속도 1 m/s 명령 성능 비교.

Fig. 8. Performance comparison of the original, baseline, and proposed controllers under a commanded lateral velocity of 1 m/s in a ship motion environment.

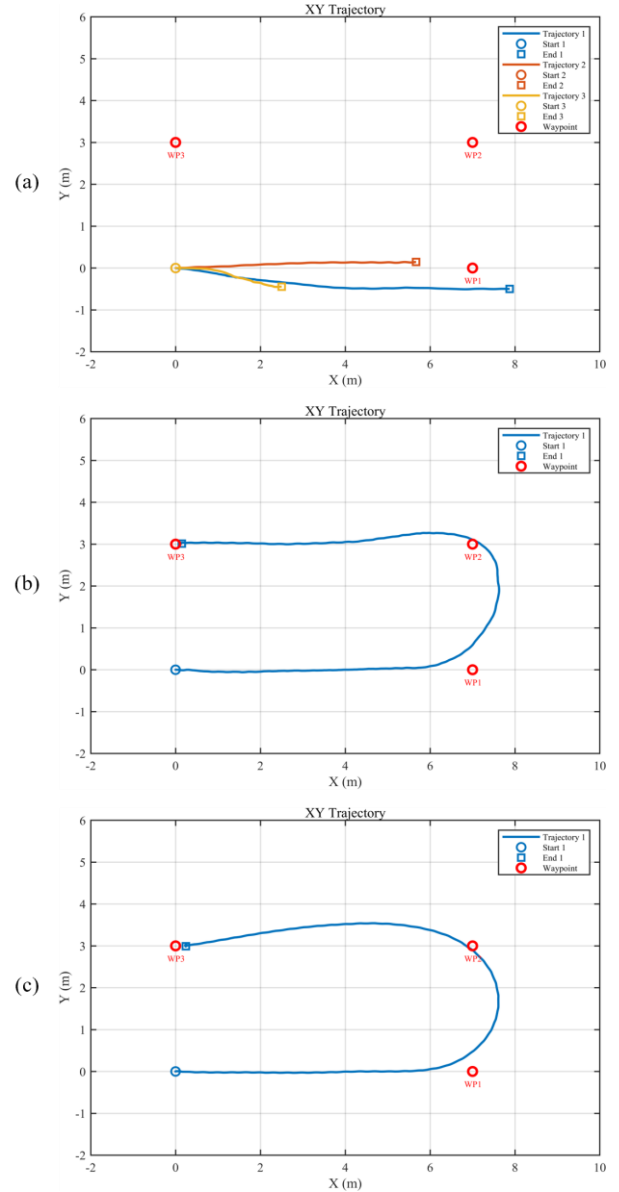


그림 9. 선박 모사 플랫폼에서 (a) original. (b) baseline. (c) proposed 제어기의 경유점 추종 성능 비교.

Fig. 9. Waypoint tracking performance comparison of (a) original, (b) baseline, and (c) proposed controllers on a ship motion simulation platform.

수 있다. 그림 8은 1m/s 종방향 속도 명령 하에 로봇의 몸체 속도를 비교한 그래프이다. 그림 8에서 볼 수 있듯이, Original 제어기는 선박의 움직임으로 인해 몸체가 지면과 충돌함을 확인할 수 있다. 반면 baseline, proposed 제어기는 안정적으로 작동함을 확인할 수 있다. 또한 proposed 제어기는 baseline 제어기에 비해 속도 프로파일에서 진동이 줄어들었다.

2. 경유점 추종 제어 성능 평가

제안된 제어기의 경유점 추종 성능을 검증하기 위해, 그림 5-(b)와 같이 진폭 15°, 주기 15초로 종방향 회전 운동을 하는 선박 운동 모사 플랫폼 환경에서 original, baseline, proposed 제어기를 적용하였으며, Pure Pursuit 기반의 경유점 추종 실험을 진행하여 제어기 간 성능을 비교하였다.

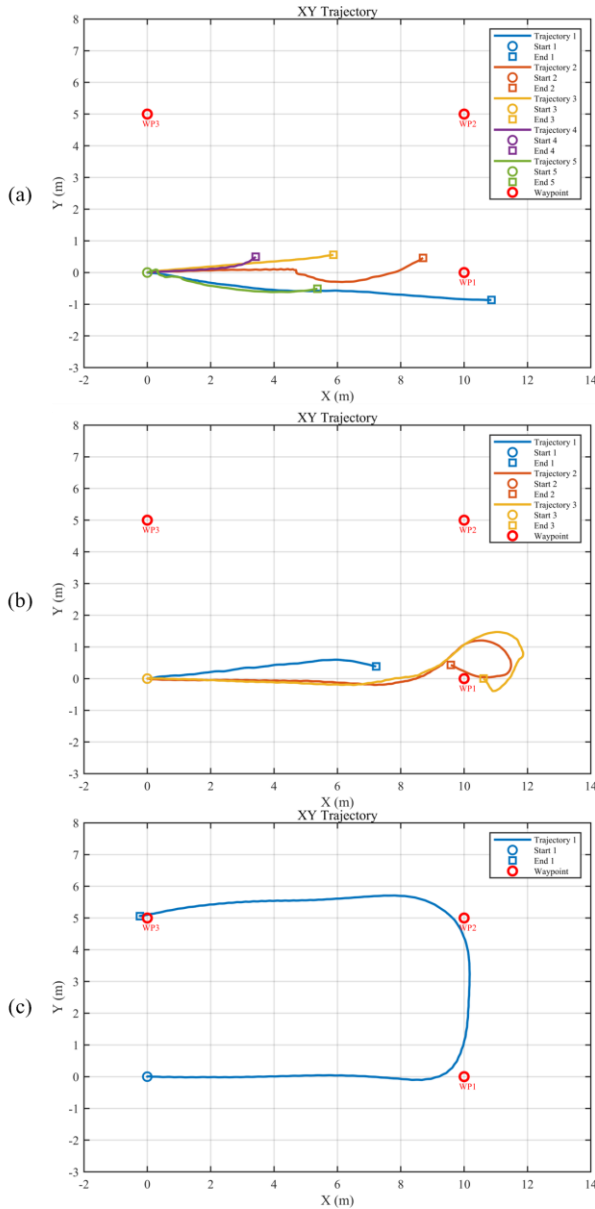


그림 10. 선박 모사 플랫폼에서 (a) original, (b) baseline, (c) proposed 제어기의 경유점 추종 성능 비교

Fig. 10. Waypoint tracking performance comparison of (a) original, (b) baseline, and (c) proposed controllers on a ship motion simulation platform.

표 4. Baseline, proposed 제어기의 경유점 추종 성능 비교.

Table 4. Comparison of waypoint tracking performance for the baseline and proposed controllers.

	RMSE (m)	Mean (m)	Max Distance (m)
Baseline	1.6642	1.0936	3.0586
Proposed	1.5018	0.9503	3.0330

그림 9는 경유점을 (7, 0), (7, 3), (0, 3)으로 설정하여 실험을 진행한 결과이다. 그림 9(a)에서 보듯이 original 제어기의 경우 지면의 움직임에 적응하지 못하여 첫 번째 경유점에 도달하지 못하고 넘어지는 결과를 보인다. 반면 baseline과 proposed 제어기의 경우 그림 9(b)에서 보듯이 안정적으로

경유점에 도달하며 마지막 도착 지점에 도달하는 것을 확인할 수 있다. 또한 표 4와 같이 baseline, proposed 제어기를 비교한 결과 baseline 제어기에 비해 본 논문에서 개발한 proposed 제어기의 경우 경로와의 오차를 통해 계산한 RMSE (Root Mean Square Error)가 약 9.8% 감소하여 더욱 정확한 경유점 추종 성능을 확보함을 확인할 수 있다.

회전 운동을 하는 지면에서는 회전 중심으로부터의 거리가 증가할수록 표면의 선형 속도가 커지게 된다. 따라서 지면 위의 물체는 회전 중심에서 멀어질수록 더욱 큰 외란을 경험하게 되며, 이는 제어 난이도를 증가시키는 요인으로 작용한다. 이에 대한 성능을 검증하기 위해 그림 10과 같이 경유점을 (10, 0), (10, 5), (0, 5)로 설정하여 더욱 큰 외란을 받는 환경에서의 실험을 진행하였다. 그림 10과 같이 original 제어기는 이전 실험과 동일하게 지면의 움직임에 적응하지 못하고 넘어지는 현상을 보였다. Baseline 제어기의 경우 이전 실험과 달리 경유점에 도달하지 못하고 넘어지는 현상을 보이며 불안정한 결과를 보인다. 이는 회전 중심에서 멀어질수록 더욱 큰 외란을 받게되어 제어 난이도가 상승하며, 정적인 비정형 환경에서 학습된 제어기는 이러한 동적 외란을 경험하지 못하였기 때문에 제어 성능 저하가 발생함을 확인하였다. 반면 proposed 제어기의 경우 이전 실험과 같이 안정적으로 경유점에 도달하는 것을 확인할 수 있다.

V. 결론

본 연구에서는 요동하는 선박에서 안정적으로 사족보행 로봇의 보행 제어를 위해 선박 운동 모사 플랫폼을 구현하였고, 해당 환경에서 PPO 강화 학습 알고리즘을 활용하여 보행 제어기를 개발하였다. 또한 평지에서 학습한 제어기, 비정형 환경에서 학습한 제어기들과 비교하여 개발한 제어기의 성능을 검증함으로써 불안정한 움직이는 지면에서도 견고한 동작이 가능함을 입증하였다. 본 연구를 통해 선박과 같은 동적 환경에서 사족 로봇을 운용하기 위해서는 기존 연구와 달리 지면의 동적 운동을 고려할 필요가 있음을 확인하였다. 본 연구에서는 이를 고려하여 강화학습 알고리즘의 보상 함수에 추가하여 동적 환경에서 안정적으로 작동하는 사족 로봇 보행 제어기를 개발하였다. 향후 연구에서는 실제 로봇에 적용하여 시뮬레이션을 통해 학습한 제어기의 현실 적용 가능성을 검증하고, 시각 센서를 통합하여 동적 환경에서의 보행 능력을 향상시킬 계획이다.

REFERENCES

- [1] F. Bonnin-Pascual, and A. Ortiz, "On the use of robots and vision technologies for the inspection of vessels: A survey on recent advances," *Ocean Engineering*, vol. 190, p. 106420, 2019. doi: 10.1016/j.oceaneng.2019.106420.
- [2] M. Ramezani, M. Brandão, B. Casseau, I. Havoutis, and M. Fallon, "Legged robots for autonomous inspection and monitoring of offshore assets," *Proc. of the Offshore Technology Conference*, pp. 1-10, 2020. doi: <https://doi.org/10.4043/30694-MS>
- [3] B. Katz, J. Di Carlo, and S. Kim, "Mini cheetah: A platform for pushing the limits of dynamic quadruped control," *Proc. of the*

- IEEE International Conference on Robotics and Automation*, pp. 6295–6301, 2019.
- [4] P. M. Wensing, M. Posa, Y. Hu, A. Escande, N. Mansard, and A. Del Prete, “Optimization-based control for dynamic legged robots,” *IEEE Transactions on Robotics*, vol. 40, pp. 43–63, 2024.
doi: <https://doi.org/10.1109/TRO.2023.3324580>
- [5] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, “Learning agile and dynamic motor skills for legged robots,” *Science Robotics*, vol. 4, no. 26, eaau5872, 2019.
doi: <https://doi.org/10.1126/scirobotics.aau5872>
- [6] D. J. Kim, J. Park, and I. Lee, “Symmetric parkour learning for robust quadruped robot control on challenging terrain,” *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 31, no. 2, pp. 106–113, Feb. 2025.
doi: 10.5302/J.ICROS.2025.24.0270
- [7] Y.-L. Cho and B.-K. Cho, “Hybrid control of wheeled quadruped robot for enhanced mobility and efficiency in diverse terrains,” *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 8, pp. 863–870, Aug. 2024.
doi: 10.5302/J.ICROS.2024.24.0055.
- [8] N. Rudin, J. Allshire, S. Hwang, R. Farshidian, and M. Hutter, “Learning to walk in minutes using massively parallel deep reinforcement learning,” *Proc. of the Conference on Robot Learning*, vol. 164, pp. 91–100, 2021.
- [9] G. B. Margolis and P. Agrawal, “Walk these ways: Tuning robot control for generalization with multiplicity of behavior,” *Proc. of the Conference on Robot Learning*, 2023, pp. 22–31.
- [10] S. Lyu, X. Lang, H. Zhao, H. Zhang, P. Ding, and D. Wang, “RL2AC: Reinforcement learning-based rapid online adaptive control for legged robot robust locomotion,” *Proc. of Robotics: Science and Systems*, 2024.
doi: <https://doi.org/10.15607/RSS.2024.XX.060>
- [11] A. Kumar, Z. Fu, D. Pathak, and J. Malik, “RMA: Rapid motor adaptation for legged robots,” *Proc. Robotics: Science and Systems*, Virtual, Jul. 2021.
doi: 10.15607/RSS.2021.XVII.011.
- [12] Z. Yan, H. Ji, and Q. Chang, “Energy consumption minimization of quadruped robot based on reinforcement learning of DDPG algorithm,” *Actuators*, vol. 13, no. 1, p. 18, 2024.
doi: 10.3390/act13010018.
- [13] H. Sun, T. Fu, Y. Ling, and C. He, “Adaptive quadruped balance control for dynamic environments using maximum-entropy reinforcement learning,” *Sensors*, vol. 21, no. 17, p. 5907, 2021.
doi: 10.3390/s21175907.
- [14] M. Yoon, H. Shin, J. Jeong, and S.-E. Yoon, “Learning-based adaptive control of quadruped robots for active stabilization on moving platforms,” *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 701–708, 2024.
doi: <https://doi.org/10.1109/IROS58592.2024.10802150>
- [15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [16] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” *Proc. 32nd International Conference on Machine Learning*, vol. 37, pp. 1889–1897, 2015.
- [17] C. K. Liu and D. Negru, “The role of physics-based simulators in robotics,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 35–58, 2021.
doi: 10.1146/annurev-control-072220-093055
- [18] N. Koenig and A. Howard, “Design and use paradigms for Gazebo, an open-source multi-robot simulator,” *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, pp. 2149–2154, 2004.
doi: 10.1109/IROS.2004.1389727.
- [19] E. Coumans and Y. Bai, “PyBullet: A Python module for physics simulation for games, robotics, and machine learning,” 2016–2021.
- [20] C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. Yuan, R. Singh, Y. Guo, H. Mazhar, A. Mandlekar, B. Babich, G. State, M. Hutter, and A. Garg, “Orbit: A unified simulation framework for interactive robot learning environments,” *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3740–3747, 2023.
doi: 10.1109/LRA.2023.3270034
- [21] T. Wang, R. Skulstad, F. T. Holmeset, K. H. Halse, H. P. Hildre, and H. Zhang, “Full-scale experimental research on wave shielding effect of RV Gunnerus for offshore operations,” *Ocean Engineering*, vol. 320, 120189, 2025.
doi: <https://doi.org/10.1016/j.oceaneng.2024.120189>
- [22] J. Zhang, C. C. Cheah, and S. H. Collins, “Experimental comparison of torque control methods on an ankle exoskeleton during human walking,” *Proc. IEEE International Conference on Robotics and Automation*, Seattle, WA, USA, 2015, pp. 5584–5589.
doi: 10.1109/ICRA.2015.7139980.
- [23] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” *Proc. IEEE/RSJ International Conference on Intelligent Robots and System*, Vancouver, Canada, 2017, pp. 23–30.
doi: 10.1109/IROS.2017.8202133
- [24] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” *Proc. 35th Int. Conf. on Machine Learning*, vol. 80, pp. 1861–1870, 2018.



구 본 석

2024년 충남대학교 선박해양공학과(공학사). 2024년~현재 충남대학교 자율운항시스템공학과 석사과정 재학 중. 관심 분야는 강화학습, 보행제어, 딥러닝, 객체 인식.



정 종 대

2008년 한양대학교 토목공학과(공학사). 2010년 KAIST 건설및환경공학과(공학석사). 2015년 KAIST 건설및환경공학과(공학박사). 2016년~2022년 선박해양플랜트연구소 선임연구원. 2022년~현재 충남대학교 자율운항시스템공학과 조교수.

관심분야는 해양로봇, 자율항법, 군집시스템, 협력제어.