JPNT 14(4), 421-429 (2025)
https://doi.org/10.11003/JPNT.2025.14.4.421

JPNT Journal of Positioning, Navigation, and Timing

강화학습 보상 설계를 통한 드론 함상 자율 착륙 강인성 개선

최보경[1], 정우주[1], 변민수[1], 정소영[1], 송진우[1], 김용훈[2†]

Reinforcement Learning Reward Design for Robust Autonomous Shipboard Drone Landing

Bokyung Choi[1], Woo Joo Jung[1], Min Su Byeon[1], So Yeong Jung[1], Jin Woo Song[1], Yong Hun Kim[2†]

[1]Department of Artificial Intelligence and Robotics and Convergence Engineering for Intelligent Drone, Artificial Intelligence and Robotics Institute (AIRI), Sejong University, Seoul 05006, Korea
[2]Department of Artificial Intelligence and Robotics, and AIRI, Sejong University, Seoul 05006, Korea

ABSTRACT

The autonomous landing of multirotor UAVs on ship decks is challenging due to wave-induced deck motion, degraded visibility, Global Positioning System (GPS) interference, and communication uncertainties. To address this, a reward design framework based on reinforcement learning for vertical drone landing on a heaving shipborne platform using the Deep Deterministic Policy Gradient (DDPG) algorithm is developed in this study. The training environment combines a simplified vertical UAV dynamics model with wave profiles generated from the JONSWAP spectrum to enable randomized and realistic heave motion in each episode. To enhance training stability and policy robustness, we introduce a distance-based reward, a strict terminal penalty for failed landings, and hyperparameter scaling consistent with vehicle dynamics. MATLAB simulation results show that the proposed reward design achieves stable policy convergence and landing performance across diverse wave conditions. These results demonstrate the proposed reward model effectively improves the learning efficiency and robustness of autonomous shipboard landing systems.

Keywords: autonomous landing, DDPG, deep RL, drone
주요어: 자율 착륙, 심층 결정적 정책 경사, 심층 강화학습, 드론

1. 서론

최근 무인항공기는 통신 중계, 표적 획득, 전술 지원 등 다양한 군사 임무의 핵심 플랫폼으로 활용되고 있다. 이 중 회전익 드론은 수직 이착륙과 정지비행이 가능해 해상 감시, 정찰 및 물자 수송과 같은 해상 작전에 특히 적합하며, 임무 장비 교체가 용이해 범용성 또한 높다 (Lee 2020). 하지만, 해상 환경에서는 파도로 인해 착륙 플랫폼이 지속적으로 상하로 움직이는 동적 특성을 가지며, Global Positioning System (GPS) 교란, 가시성 저하, 통신 불확실성과 같은 요인으로 인해 조종자가 수동으로 착륙하기 어려운 경우가 빈번하다. 따라서 좁은 갑판 위 드론을 안정적이고 자율적으로 착륙시키는 기술은 해상 작전의 지속성을 위한 핵심 요소이다. 이 때문에 드론이 외부 조종 없이 플랫폼의 상태를 스스로 인식하고 안정적으로 착륙할 수 있도록 하는 자율 착륙 기술에 대한 다양한 연구가 진행되고 있다 (Subamanian et al. 2023).

기존의 방법 중 Proportional-Integral-Derivative (PID) 제어기를 활용한 착륙 시스템은 고정된 제어 이득에 의존하기 때문에 파도와 같이 예측할 수 없는 환경 변화에 실시간 대응하기 어렵다는 한계가 존재한다 (Talha et al. 2019). 또한, 지도학습 기반 모델 역

Received Nov 20, 2025 **Revised** Nov 28, 2025 **Accepted** Dec 04, 2025

†**Corresponding Author** E-mail: yhunkim@sejong.ac.kr

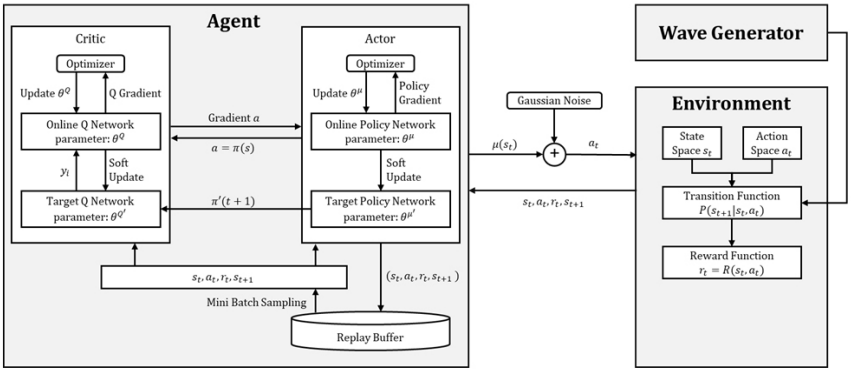


Fig. 1. Overview of a reinforcement learning-based autonomous landing system.

시 드론의 6 자유도 운동 특성에 따른 모든 행동을 사전에 정의하거나 학습을 위한 방대한 실험 데이터를 확보하기가 어려워 개발 비용이 크게 증가하는 한계가 있다. 즉, 두 접근법 모두 실제 해상과 같은 동적 환경에 적용이 제한적이라는 문제가 있다.

이러한 한계를 극복하기 위해 다양한 상황에서의 경험으로 에이전트가 스스로 학습할 수 있는 심층 강화학습(Deep Reinforcement Learning, DRL) (François-Lavet et al. 2018, Wang et al. 2024) 기반 제어 기법이 주목받고 있다 (Amendola et al. 2024). DRL은 환경과의 상호작용을 통해 모델링 오차나 예측 불확실성이 존재하는 상황에서도 적응적인 정책(policy)을 학습할 수 있으며, 최근 연구에서는 바람, 플랫폼 속도 변화 등 비정형 외란이 존재하는 상황에서도 높은 제어 성능을 보이고 있다. Rodriguez-Ramos et al. (2019)의 연구에서는 fiducial marker를 통해 얻은 위치, 자세 정보를 추정하여 이를 Deep Deterministic Policy Gradient (DDPG)의 상태 입력으로 사용해 직선 및 8자 형태로 움직이는 플랫폼에 착륙을 시도하였고, 다양한 속도, 궤적 시나리오에서 정책을 학습하였다. 그러나 연속 상태 및 액션 공간 문제에 집중하기 위해 수직축에 대한 제어는 상수 속도 참조 방식을 사용하였다. Xie et al. (2020)의 연구에서는 DDPG를 사용하였으나 수직 하강 속도는 사전 정의된 규칙을 따르고 수평 제어만 DRL이 담당하도록 하였으며, 그 결과 PID 대비 약 10% 높은 착륙 성공률을 달성하였다.

기존 연구들은 주로 강화학습을 수평 운동에만 적용하고, 수직축은 PID나 고정 하강 속도에 의존하는 경우 (Wu et al. 2022)가 많아 파도에 따른 플랫폼의 상하 운동(heave motion)을 정책 수준에서 다루지 못하는 한계가 있다. 그러나 해양 플랫폼은 비정규적인 상하 움직임에 의한 외란이 지속적으로 발생하기 때문에, 수직축의 환경 변화를 직접 반영한 학습 기반 착륙 전략이 필요하다. 이에 본 연구는 플랫폼의 상하 운동을 포함한 3차원 환경에서 드론이 수직 착륙 정책을 DDPG로 학습하도록 설계하였으며, 제안된 보상 구조를 통해 다양한 파도 조건에서도 안정적인 착륙 동작을 학습함을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 심층 강화학습과 DDPG의 기본 이론을 정리하고, 3장에서는 이를 기반으로 한 자율 착륙 정책 설계 방법을 제시한다. 4장에서는 시뮬레이션 환경 구축과 실험 설정을 설명하며, 5장에서는 제안된 기법의 성능을 다양한 시나리오에서 검증한다. 마지막으로 6장에서는 연구 결과를 요약하고 향후 연구 방향을 제안한다.

2. 이론적 배경

2.1 심층 강화학습

강화학습은 에이전트가 환경과 상호작용하며 누적 보상을 최대화하는 방향으로 최적의 행동 정책을 학습하는 방법론이다. 그러나 전통적인 RL은 상태(state) 공간과 행동(action) 공간이 커지거나 연속적으로 확장될 경우 정책 또는 가치 함수(value function)를 테이블 형태로 표현하기 어려워 실제 제어 문제에 적용하는 데 한계가 있다. 이러한 문제를 해결하기 위해 심층 신경망을 결합한 DRL이 도입되었고, 신경망을 활용해 정책 또는 가치 함수를 근사함으로써 고차원 연속 제어 문제에서도 안정적인 학습이 가능해졌다 (François-Lavet et al. 2018, Wang et al. 2024).

2.2 DDPG 알고리즘

DDPG는 actor-critic 구조의 off-policy DRL 알고리즘으로, actor는 현재 상태에서 연속적인 제어 입력을 생성하고, critic은 해당 상태-행동 쌍의 Q-value를 평가한다 (Lillicrap et al. 2016). 학습 안정성을 위해 타겟 네트워크(target network)를 사용하며, 경험 재현 메모리(replay buffer)를 통해 시간적 상관성이 제거된 경험을 샘플링하여 학습한다. 또한, 정책에 탐색 노이즈를 주입하여 연속 공간에서의 행동 탐색을 수행한다. DDPG는 로봇 제어, 드론 자세 안정화, 경로 추종 등 연속 제어가 요구되는 분야에서 널리 사용되고 있으며, 본 연구에서는 해상 환경에서의 자율 착륙과 같이 예측 불가능한 외란에 대응하면서도 안정적인 수렴이 필요한 수직 제어 문제에 적용하고자 한다.

3. DDPG 기반 착륙 유도 구조 설계

본 연구에서 제안하는 DDPG 기반 착륙 유도 구조는 Fig. 1과 같이 크게 세 가지로 구성된다.

- 에이전트(agent): 상태와 행동을 통해 환경과 상호작용하며, DDPG 알고리즘으로 학습한다.
- 환경(environment): 드론의 수직 운동 모델과 플랫폼의 상하 운동을 포함하며, 에이전트의 행동에 따른 보상과 새로운 상태를 반환한다.
- 파도 생성기(wave generator): 실제 해상 환경을 모사하기 위해 유의파고 및 파 주기를 기반으로 랜덤한 파형을 생성한다.

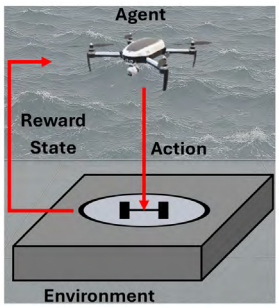


Fig. 2. Example scenario of reinforcement learning-based drone landing.

3.1 드론의 수직 운동 모델링

일반적으로 드론의 운동은 x, y, z축에 대한 병진 운동과 roll, pitch, yaw 회전 운동, 즉 6 자유도 운동을 하며, 이는 복잡한 비선형 동역학 방정식으로 나타내어진다. 본 연구에서는 드론이 수직 방향으로 움직이는 플랫폼에 안전하게 착륙하는 문제에 집중하여 전체 6 자유도 운동 중 수직축 운동을 분리하여 네 개의 모터에서 발생하는 총 추력을 하나의 추력으로 모델링하였다.

드론의 수직 운동은 추력과 중력(g)의 상호작용으로 결정되며, 호버링(hovering) 상태에서는 총 추력 T가 무게 mg와 평형을 이루어 $T \approx mg$가 된다. 추력이 중력보다 크면 상승하고 작으면 하강하며, 실제로는 네 개의 모터 회전 속도를 동일하게 높이거나 낮춤으로써 수직 추력을 변화시켜 고도를 제어한다. 이 구조를 기반으로 한 수직 방향 동역학은 Eq. (1)과 같이 단순화된 형태로 표현된다.

$$m\ddot{z} = T - mg \tag{1}$$

여기서 m은 드론의 질량, $\ddot{z}$는 z축 가속도이다. 추력을 에이전트가 출력하는 제어입력 u로 표현하고, 정리하면 Eq. (2)와 같다.

$$\ddot{z} = \frac{1}{m}(u - mg) \tag{2}$$

드론의 질량은 1.5 kg, 중력 가속도는 $g = 9.81\left[\frac{m}{s^2}\right]$로 정의한다.

3.2 상태, 행동 공간 및 제어 입력 정의

강화학습 기반 제어 문제를 정의하기 위해서는 에이전트인 드론의 상태와 행동을 명확히 정의해야 한다. 드론이 상하로 움직이는 플랫폼 위에 안정적으로 착륙하는 문제를 다루기 위해, 수직 운동과 플랫폼의 동역학적 특성을 반영하여 Fig. 2와 같이 시나리오 구성을 위한 기본 조건을 가정하고, 상태와 행동 공간을 정의하였다.

3.2.1 상태 및 행동 공간 정의

드론의 수직 착륙 문제를 강화학습 환경으로 정의하기 위해, Eq. (3)과 같이 상태벡터를 드론과 플랫폼의 위치 및 속도 정보를 포함한 4차원 벡터로 정의하였다.

$$s_t = \left[z_d, \dot{z}_d, z_p, \dot{z}_p\right]^T \tag{3}$$

이 벡터는 드론의 고도(z_d), 수직 속도($\dot{z}_d$), 플랫폼의 고도(z_d), 그리고 플랫폼의 수직 속도($\dot{z}_p$)를 벡터의 요소로 갖는다. 이는 에이전트가 플랫폼의 위치와 움직임까지 고려하여 착륙 전략을 학습할 수 있도록 하기 위함이다. 시뮬레이션에서는 상태벡터가 네 가지 요소로 구성되지만, 실제 환경에서는 이러한 절대적 상태 정보를 직접 얻기 어렵다. 따라서 실제 적용 시에는 Light Detection and Ranging (LiDAR), Radio Detection And Ranging (RADAR), Depth Camera 등과 같은 하향식 거리 측정 센서로부터 추정된 상대 위치 및 상대 속도 정보로 상태벡터를 대체하여 구성할 수 있다.

3.2.2 행동 공간 및 제어 입력 변환

에이전트의 행동 a_t는 드론의 z축 추력을 조절하는 스칼라 값이다. 행동 공간은 [−1, 1] 범위의 1차원 연속 제어 입력으로 구성되고, 에이전트가 선택한 값은 Eq. (4)를 사용해 실제 물리 추력으로 변환된다.

$$u = mg + u_{max}a_t \tag{4}$$

여기서 m은 1.5 kg 드론의 질량을, u_{max}는 최대 추력 15 N을 의미한다. 이 변환 식은 에이전트가 출력한 행동 값($a_t \in [-1,1]$)을 물리적 제약 조건인 $0 \le T \le u_{max}$ 실제 추력 T로 매핑하는 역할을 한다. 결과적으로 에이전트는 이 스칼라 값을 통해 드론의 추력을 제어하고 최적 정책을 학습한다.

3.3 보상 함수 설계

강화학습 문제에서 보상 함수는 에이전트가 목표를 효과적으로 학습하도록 유도하는 핵심 요소이고 특히 드론의 동적 플랫폼 착륙 연구에서 보상 함수의 설계는 매우 중요한 요소이다. 본 연구는 '플랫폼과 드론의 거리 감소'와 '충돌 없는 안전한 착륙'이라는 두 가지 목적을 동시에 달성해야 한다.

그러나 이 두 목표는 학습 과정에서 상충하는 목표를 가진다. 플랫폼과의 거리를 빠르게 줄이는 고속 하강과 충돌을 방지하기 위한 상대 속도 감소는 물리적으로 상반되는 행동을 요구하므로, 이를 동시에 보상으로 부여하는 경우 에이전트가 일관된 정책을 학습하기 어렵다. 예를 들어, 플랫폼이 상승할 때 드론이 위치 오차 보상을 위해 단순히 거리 오차를 줄여 하강 속도를 높이면, 상대 속도를 고려하지 못해 플랫폼과 충돌할 위험이 커진다. 이처럼 상반된 목적을 동시에 보상으로 부여하면 정규화되지 않은 서로 다른 보상 함수로 인해 학습이 불안정해진다. 이러한 복잡성

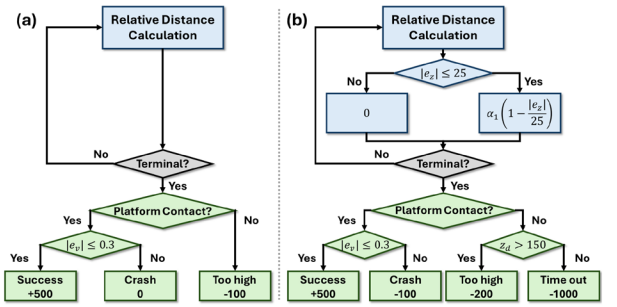


Fig. 3. Comparison of reward structure (a) Baseline reward (b) Proposed reward.

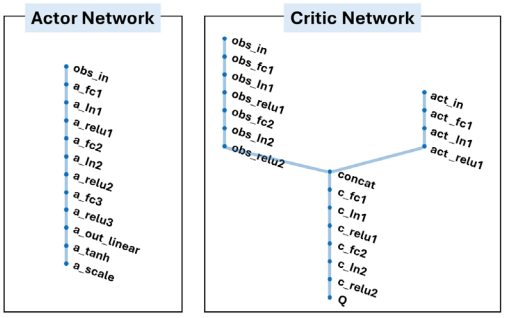


Fig. 4. Actor network and critic network architecture.

을 줄이고 학습 목표를 명확히 하고자, 불필요한 시간 페널티나 속도 관련 항을 제거한 보상 함수를 제안한다. 연속적인 보상으로 드론과 플랫폼 간의 상대 위치 오차항만을 사용하여 에이전트가 플랫폼에 접근하도록 한다.

베이스라인과 제안하는 보상 구조를 Fig. 3에 나타내었다. Fig. 3a는 기존 방식의 터미널 보상 중심 구조를, Fig. 3b는 연구에서 제안하는 거리 기반 연속보상과 터미널 보상을 결합한 구조이다.

Fig. 3a는 단순하게 에피소드 종료 시점에서 '안정적 착륙 성공'과 '충돌 실패'라는 터미널 조건에 각각 대비되는 희소 보상을 부여하였다. 희소 보상(sparse reward)이란 에이전트가 학습 중 대부분 단계에서 보상을 받지 못하고, 목표 달성 또는 에피소드 종료 시점에만 보상이 주어지는 것을 의미한다. 본 연구에서 제안된 보상 구조인 Fig. 3b는 상대 거리에 대한 보상을 스텝마다 부여하여 안정적인 착륙에 성공할 수 있도록 하였다. 이러한 보상 설계를 통해, 에이전트는 상충하는 중간 신호에 매몰되지 않고, 단순히 플랫폼에 근접하는 것을 넘어 최종적으로 충돌을 회피하여 안전하게 착륙하는 정책을 학습할 수 있었다.

드론과 플랫폼 간의 상대고도 오차가 줄어들수록 거리에 대한 보상($r_{distance}$)이 커지도록 보상을 Eqs. (5, 6)과 같이 정의하였다.

$$e_z = z_d - z_p \tag{5}$$

$$r_{distance} = \begin{cases} \alpha_1\left(1-\dfrac{|e_z|}{25}\right), & \text{if } e_z \le 25 \\ 0, & \text{if } e_z > 25 \end{cases} \tag{6}$$

여기서 e_z는 드론과 플랫폼 사이의 상대 거리, α_1은 보상함수의 사용자 정의 가중치이다. 이 식의 값은 드론이 플랫폼에 가까워질수록 증가하며, 두 객체의 고도 차이가 0일 때 최대값을 갖는다. 에피소드 종료 시점에는 착륙 성능을 보다 명확하게 반영하기 위해 Eq. (7)과 같이 속도 오차를 구하고 Eq. (8)의 터미널 보상($r_{terminal}$)을 부여한다.

$$e_v = \dot{z_d} - \dot{z_p} \tag{7}$$

$$r_{terminal} = \begin{cases} R_{soft} = 500, & \text{if } |e_z| < 0.1 \text{ and } |e_v| \le 0.3 \\ R_{crash} = -100, & \text{if } |e_z| < 0.1 \\ R_{toohigh} = -200, & \text{if } z_d > 150 \\ R_{timeout} = -1000, & \text{if step} > 4500 \end{cases} \tag{8}$$

에피소드 종료 조건은 착륙 성공(R_{soft}), 충돌(R_{crash}), 고도 초과($R_{toohigh}$) 그리고 최대 시뮬레이션 시간 초과($R_{timeout}$)로 정의하였다. 드론이 플랫폼에 성공적으로 착륙했을 경우 +500의 보상을 부여한다. 착륙 성공의 조건은 드론과 플랫폼 간의 상대고도 오차(e_z)가 0.1 m 미만이고, 상대 속도 오차(e_v)가 0.3 m/s 이하일 때로 설정하였다. 드론이 플랫폼에 착륙하였으나 빠른 속도로 착륙해 충돌하는 경우에는 불안정한 착륙으로 간주하고 -100의 페널티를 부여한다. 또, 드론이 150 m 이상으로 상승하는 경우에 -200의 페널티를 주고 해당 에피소드는 종료된다. 마지막으로, 최대 스텝을 넘어가는 경우 -1000의 큰 페널티를 부여한다.

3.4 네트워크 구조

DDPG 에이전트는 Actor-Critic 구조의 심층 신경망으로 설계되었으며, 움직이는 착륙 지점과 안전 제약을 동시에 고려하기 위해 [256, 256, 128] 유닛의 다층 퍼셉트론 구조로 드론 착륙 문제의 복잡한 비선형 동역학을 충분히 표현할 수 있도록 구성하였다.

Actor 네트워크는 상태로부터 고도, 속도, 플랫폼 위치 간의 관계를 계층적으로 표현하도록 설계되었으며, Fully Connected (FC) Layer와 ReLU 활성층을 반복적으로 구성하여 비선형 제어 정책을 근사한다. 최종 출력단에는 tanh 함수를 적용해 행동 값을 [-1, 1] 범위로 제한하고, 이후 scaling layer를 통해 드론의 실제 추력 한계 범위로 선형 변환함으로써 학습 과정에서 비정상적 출력으로 인한 불안정성을 방지하였다.

Critic 네트워크는 상태와 행동 입력을 초기 단계에서 분리하여 각각의 의미적 차이를 추출한 후, 두 경로에서 추출된 특징을 결합하고 FC Layer를 통과시켜 Q-value를 근사한다. 이러한 구조는 상태-행동 상호작용을 비선형적으로 평가할 수 있어 정책 개선의 안정성과 수렴성을 향상한다. 또한, 모든 은닉층에 Layer 정규화를 적용하여, 경험 리플레이로 인한 샘플 비독립성 문제를 완화하고 학습의 안정성을 높였다. 제안된 네트워크 구조는 Fig. 4에 나타내었다.

4. 시뮬레이션 환경 및 해상 플랫폼 구성

앞서 제시한 드론의 수직 운동 모델과 DDPG 알고리즘 기반 착륙 시스템을 검증하기 위해, MATLAB 기반 시뮬레이션 환경을

Table 1. WMO sea state code.

Sea state code	Wave height [m]	Characteristics
0	0	Calm (glassy)
1	0 to 0.1	Calm (rippled)
2	0.1 to 0.5	Smooth (wavelets)
3	0.5 to 1.25	Slight
4	1.25 to 2.5	Moderate
5	2.5 to 4	Rough
6	4 to 6	Very rough
7	6 to 9	High
8	9 to 14	Very high
9	Over 14	Phenomenal

구축하였다.

4.1 MATLAB 시뮬레이션 환경 구성

각 에피소드는 드론이 초기 고도 25 m에서 중력과 추력이 평형을 이루는 호버링 상태에서 시작한다. 시뮬레이션의 타임 스텝(Δt)은 0.01초이며, 에피소드 당 최대 길이는 4500 스텝으로, 총 45초로 제한된다. 이는 드론이 무한히 상승하거나 추락하는 비정상적인 궤적을 탐색하는 것을 방지하고, 제한된 시간 내에 착륙 임무 완수 여부를 평가할 수 있도록 하기 위함이다. 초기 고도 25 m는 에이전트가 목표 플랫폼에 접근하는 과정에서 안정적인 제어 정책을 학습하기에 충분한 시간을 확보할 수 있도록 설정하였다.

4.2 Heaving 플랫폼 모델링

플랫폼의 상하 운동을 재현하고자 해상 파도 시뮬레이션 환경을 구축하였다. 파도 모델로는 최근까지도 실제 파도 모사를 위한 표준으로 통용되는 JONSWAP 스펙트럼을 적용하였으며, 이는 시뮬레이션 툴에서도 널리 사용된다 (McTaggart 2012, Kim & Shin 2019, Palmer & Irani 2026). 스펙트럼은 Eq. (9)와 같이 정의되며, 각 파라미터의 의미는 (Det Norske Veritas 2011)에 정의된 것과 동일하게 사용하였다.

$$S(\omega) = A_\gamma \frac{5}{16} H_S^2 \, \omega_p^4 \, \omega^{-5} \, e^{-\frac{5}{4}\left(\frac{\omega}{\omega_p}\right)^{-4}} \gamma^{e^{-0.5\left(\frac{\omega-\omega_p}{\sigma\omega_p}\right)^2}} \tag{9}$$

파도는 서로 다른 진폭과 위상을 갖는 다수의 정현파 성분의 중첩으로 표현될 수 있어 주파수 영역에서 정의된 파도 스펙트럼 $S(\omega)$를 활용하여 Eqs. (10, 11)을 통해 시간에 따른 해수면 높이로 변환하여 시뮬레이션에서 파도를 모델링 하였다.

$$A_k(\omega) = \sqrt{2S(\omega)d\omega} \tag{10}$$

$$\eta(t) = \sum_{k=1}^{n} A_k(\omega) \cos[2\pi\omega_k t + \epsilon_k] \tag{11}$$

JONSWAP 스펙트럼의 유의파고(significant wave height, H_s)를 조정함으로써 특정 sea state 등급에 따른 파도 조건을 모사할 수 있다. 본 연구에서는 WMO Sea State 분류 기준을 바탕으로, 각 state별 최대 wave height를 유의파고 파라미터로 적용하였으며, 이를 정리하면 Table 1과 같다.

Table 1의 조건을 기반으로 Fig. 5와 같이 파고를 생성하였으며,

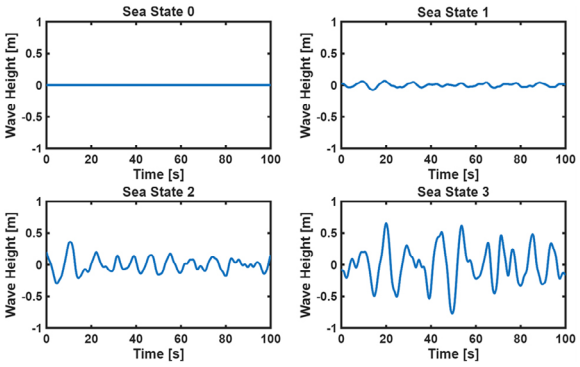


Fig. 5. Wave height according to sea state.

sea state에 따라 파도의 높이가 상이한 것을 확인할 수 있다.

정책의 일반화 성능을 확보하기 위해, 에피소드마다 유의파고 및 스펙트럼 특성은 동일하게 유지하되, 위상이 서로 다른 파형을 매번 새롭게 생성하여 학습에 적용했다. 이를 통해 에이전트가 특정 파도 패턴에 과적합되는 것을 방지하고, 동일 sea state 내에서도 다양한 수직 운동 변화에 적응하는 정책을 학습할 수 있도록 하였다.

5. 시뮬레이션 결과 및 분석

제안한 기법의 성능을 평가하기 위해, 희소 보상 환경에서 시작하여 상대 거리 기반 보상 설계와 DDPG 하이퍼파라미터 튜닝을 단계적으로 적용한 시나리오를 구성하고 분석하였다.

5.1 희소 보상 기반 학습

첫 번째 시나리오에서는 에이전트에게 희소한 보상을 제공하는 베이스라인 환경을 구성하였다. 에이전트는 에피소드 종료 시점에만 보상을 부여받으며, 안전 착륙 시 +500, 플랫폼에 충돌할 경우 0, 착륙에 실패한 경우 −100의 보상을 받도록 설계하였다.

그 결과 Fig. 6a와 같이 평균 보상이 0 근처에서 변동하지 않는 것을 확인할 수 있다. 이는 DDPG가 결정론적 정책이기 때문에 보상이 거의 없는 환경에서 actor, critic 네트워크가 더 이상 업데이트되지 않는 정책 데드락(deadlock) 현상이 발생한 것으로 해석된다 (Matheron et al. 2020). 초기 학습 단계에서 유의미한 보상을 경험하기 전, Actor의 출력은 누적되는 그래디언트에 의해 한쪽으로 점차 치우치며 포화(saturation)되고, 이후 Critic은 이러한 잘못된 정책의 가치를 그대로 학습하게 된다. 따라서 희소 보상을 받은 Q 함수는 거의 일정한 값만을 갖기 때문에, Actor 업데이트에 필요한 Critic의 기울기($\Delta_a Q$)가 사라지며 학습이 사실상 정지된다. 이 상태에서는 탐색 노이즈에 의해 우연히 보상을 발견하더라도, 그 보상 정보가 정책으로 반영되지 못해 에이전트는 데드락에서 빠져나올 수 없다. 이는 결국 드론이 지속해서 최대 추력을 출력하는 비정상적인 행동으로 나타난다.

5.2 연속 보상 기반 학습 안정화

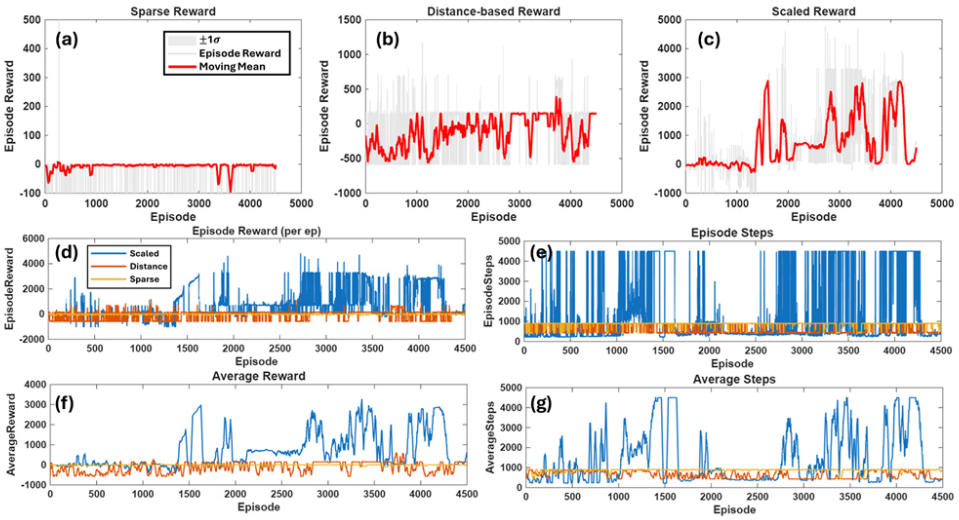


Fig. 6. Comparison of sparse, distance-based, and scaled reward schemes in DDPG training.

희소 보상 환경 문제를 완화하기 위해, 매 스텝 드론과 플랫폼 간 상대고도에 따라 Eq. (12)와 같이 보상을 부여하였다.

$$r_{dist} = \alpha_1 \left(1 - \frac{|e_z|}{z_0}\right) \tag{12}$$

여기서 r_{dist}는 매 스텝 에이전트가 받는 거리 기반 보상으로 α_1은 보상의 크기를 조절해주기 위한 가중치, e_z는 드론과 플랫폼 사이 상대고도, z_0는 드론의 초기 고도를 의미한다. 이 보상은 드론이 플랫폼 쪽으로 하강할수록 양의 보상을, 반대로 상승할수록 음의 보상을 주어 에이전트가 목표 방향으로 접근할 수 있도록 유도한다. 그러나 이 보상만으로는 하강 구간에서 얻을 수 있는 양의 보상을 탐색하기 전에 상승 방향으로 정책이 고착될 수 있어 비정상적인 정책을 학습할 가능성이 여전히 존재한다. 이러한 현상을 방지하기 위해 고도 상한을 초과할 경우 큰 페널티를 부여하고 에피소드를 종료하는 고도 제약 조건을 추가하였다. 이를 통해 비효율적인 상승 정책이 억제되고, 에이전트는 하강하며 점진적 보상을 확보하는 방향으로 정책을 학습하게 된다. 그 결과, Fig. 6b에서 1500 에피소드 이후로 에이전트가 평균 100 정도의 누적 합을 얻는 정책을 학습한 것을 확인할 수 있다. 대부분의 보상이 0 부근에 있던 Fig. 6a와 달리 양의 보상 정책을 탐험하여 비정상 상승 현상이 감소하였다.

5.3 DDPG 하이퍼파라미터 튜닝

연속 보상 구조를 Eq. (12)에서 제안한 보상식인 Eq. (6) 형태로 정교화하고, 시스템 보상 구조의 스케일과 에피소드 길이 그리고 heaving 플랫폼의 운동 특성을 고려하여 하이퍼파라미터를 튜닝한 시나리오를 구성했다. DDPG 알고리즘은 상태 및 보상 스케일에 민감하여 Q 함수가 발산하거나 포화하는 현상이 발생한다는 특징이 있다. 하지만, 상대고도 기반 보상과 터미널 착륙 보상을 추가하여 안정적으로 정책이 반영될 수 있도록 설계하였다.

에이전트의 Actor는 작은 보폭으로 정책을 갱신하고 Critic은 상대적으로 빠르게 가치를 학습하도록 하이퍼파라미터를 튜닝하였다. 또한, 타깃 네트워크의 soft-update 계수와 discount factor는 드론이 플랫폼에 접근하는 단계에서 받는 보상과 착륙 성공 여부에 대한 보상을 같은 가치 기준으로 처리할 수 있도록 설정하였다. 경험 재현 버퍼 크기와 미니배치 크기는 랜덤한 파형을 사용해 다양한 착륙 시도가 수집되는 환경을 학습할 수 있는 크기로 설정하였다. 이는 특정 파형이나 일시적인 성공 사례에 대한 과적합을 방지하고, 에이전트가 학습 과정에서 받는 보상을 통해 지속적으로 환경을 경험할 수 있도록 한다. 탐색은 초기에 상대적으로 큰 가우시안 노이즈를 사용하여 다양한 하강, 접근 전략을 시도하도록 하였으며, 학습이 진행됨에 따라 노이즈 분산을 점진적으로 감소시키는 방식으로 탐색-탐험 균형을 조절하였다.

Table 2. Training hyperparameters for the DDPG algorithm.

Hyperparameter	Value
Actor learning rate	1×10^{-4}
Critic learning rate	1×10^{-3}
Target network soft update coefficient (τ)	0.001
Discount factor (Υ)	0.99
Replay buffer size	1×10^{-6}
Mini-batch size	64
Exploration noise variance	0.6
Noise decay rate	1×10^{-5}

Table 2의 하이퍼파라미터 설정으로 학습한 결과, Fig. 6c에서 확인할 수 있듯 약 1300 에피소드 이후 에이전트가 평균 1500 수준의 누적 보상을 받는 쪽으로 학습한 것을 알 수 있다.

5.4 학습 성능 및 착륙 성능 분석

5.4.1 학습 성능 분석

보상 구조가 학습 과정에 미치는 영향을 확인하기 위해, 설계한 세 가지 시나리오에 따른 학습 성능을 분석한 결과는 Fig. 6과 같다. Figs. 6d,f는 각 보상 설계에 대한 에피소드 보상과 그 이동

Table 3. Mean time-to-land (TTL) comparison between PID and the proposed RL method.

Techniques	Mean TTL [s]	
	< 5 m	Entire interval
PID	27.830	35.402
DRL (proposed)	6.470	11.030

평균을 나타낸다. Sparse 및 Distance 기반 보상은 에피소드 보상이 0 부근의 범위에 머물며 학습이 진행되어도 평균 보상이 유의미하게 증가하지 않는다. 이는 에이전트가 성공 궤적을 거의 발견하지 못하고, 초기 정책과 유사한 행동을 반복하고 있음을 의미한다. 반면, Scaled Reward는 에피소드 수가 증가함에 따라 높은 누적 보상을 가지는 에피소드가 점차 자주 학습되고, 이동 평균 또한 상승하는 추세를 보인다. 이를 통해 큰 보상을 받는 궤적을 반복적으로 탐색하는 방향으로 정책이 개선되고 있음을 확인할 수 있다.

Figs. 6e,g는 각 보상 설계에 대한 에피소드 길이와 그 평균을 비교한 것이다. Sparse와 Distance 기반 보상에서는 에피소드 대부분이 비교적 이른 시점에 종료되어 에피소드 길이가 거의 증가하지 않는다. 이는 에이전트가 플랫폼에 충분히 접근하기 전에 상승 또는 빠른 충돌과 같은 행동으로 실패하고 있음을 나타낸다. 반면, Scaled Reward에서는 학습 후반부로 갈수록 긴 스텝이 지속되는 비율이 증가한다. 이러한 결과는 제안한 보상 구조를 통해 학습하는 에이전트가 더 오래 생존하면서 플랫폼에 접근하는 장기 전략을 탐색하도록 유도한다는 것을 의미한다.

5.4.2 착륙 성능 분석

제안한 알고리즘의 강건성을 검증하기 위해, 학습된 에이전트를 대상으로 무작위 파도 위상을 적용한 100회의 시뮬레이션을 수행하였다. 실험 결과, 제안된 보상 구조를 적용한 에이전트는 98%의 높은 착륙 성공률을 기록하였다. 반면 희소 보상을 사용한 경우에는 드론이 지속적으로 고도를 높여 착륙에 실패하였고, 거리 기반 보상만을 사용한 경우에는 플랫폼에 강하게 충돌하여 착륙에 실패했다. 이러한 높은 착륙 성공률을 바탕으로 제안한 알고리즘을 PID 기반 제어기와 비교하여 착륙 효율성을 평가하였다. 평균 착륙 소요시간(Mean Time-to-Land, TTL)을 기준 지표로 사용해, 드론과 플랫폼의 상대고도가 5 m 이하인 근접구간과 전체 비행 구간으로 나누어 분석하였다. 성능 비교결과는 Table 3과 같다.

PID 기법은 파도에 의한 위치 오차를 보정하기 위해 플랫폼의 움직임을 과도하게 추종하여 Fig. 7b와 같이 파형에 따른 상승과 접근이 반복되었고, 이로 인해 TTL이 크게 증가했다. 반면, 제안한 강화학습 기법은 거리 기반 연속 보상을 통해 플랫폼의 heaving 주기에 대한 최적 하강 타이밍을 스스로 학습함으로써 빠른 착륙 성능을 달성했다. 특히, 5 m 이하 접근 구간에서는 DRL의 TTL이 PID 대비 약 4.3배 짧아졌으며, 이는 에이전트가 실시간으로 착륙 전략을 조정했기 때문이다.

다음으로 착륙 직전 드론과 플랫폼 간 속도 관계를 분석한 결과 Fig. 8과 같이 상대 속도의 크기가 0.1 m/s 이하로 낮은 속도를 유지하며 착륙했다. 이는 드론이 플랫폼의 순간 속도에 효과적으로 동기화된 상태에서 착륙하고 있음을 보여준다. Fig. 9는 드론과 플랫폼의 착륙 직전 속도를 산점도로 표현한 것이며, $y=x$ 직선을 따라 분포하는 것을 확인할 수 있다. 이는 플랫폼의 동특성이 큰 상황에서도 제안된 기법을 적용할 경우 드론이 플랫폼과 동일한 속도로 추종하며 착륙할 수 있음을 의미한다.

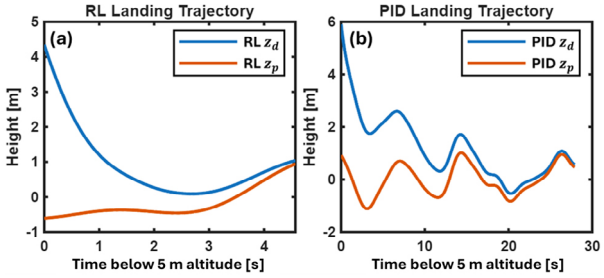


Fig. 7. Comparison of RL and PID landing trajectories below 5 m.

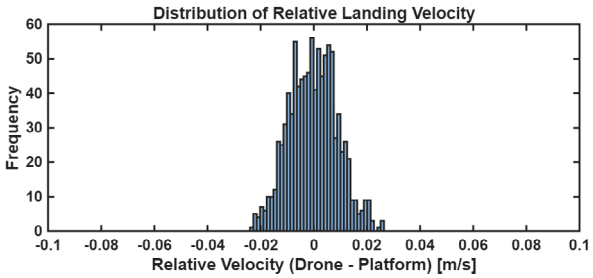


Fig. 8. Distribution of relative landing velocity.

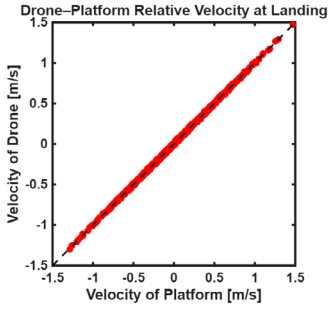


Fig. 9. Scatter plot of drone–platform velocities at landing.

6. 결론

본 연구에서는 파도로 인한 플랫폼의 상하 운동 환경에서 드론이 안정적으로 착륙할 수 있도록 DDPG 기반의 수직축 착륙 유도 시스템을 제안하고 시뮬레이션을 통해 그 유효성을 입증하였다. 상태 정보와 연속형 보상 함수를 결합한 설계를 통해, 에이전트가 불규칙한 파형 조건에서도 플랫폼에 안정적으로 착륙함을 확인하였다. 특히 보상 체계와 탐색 파라미터의 최적화를 통해 DDPG 고유의 학습 불안정성과 희소 보상 문제를 효과적으로 개선하였다. 이는 기존 연구들이 수평 제어에 집중하거나 하강률을 고정하여 문제를 단순화했던 한계를 극복한 것으로 평가할 수 있다. 결과적으로 본 연구에서 제안한 시스템은 수동 조종이 제한적인 해

상 및 함상 환경에서 드론의 자율 착륙 정확도를 향상시켜, 해상 드론 시스템의 운용 신뢰성을 높였다는 것에 의의가 있다.

향후 연구에서는 바람과 같은 외란이 존재하는 상황에서도 에이전트가 불필요한 체공 시간을 줄이고 빠른 착륙을 수행할 수 있도록, 시간에 비례하는 페널티를 도입하는 등 보상 함수를 개선할 예정이다. 이러한 개선된 보상 설계를 기반으로 본 알고리즘을 수평축 제어 기법과 결합하여 3차원 착륙 문제로 확장하고, 해상 환경에서의 알고리즘 안정성을 평가할 것이다. 또한, Twin Delayed Deep Deterministic policy gradient algorithm (TD3), Soft Actor-Critic 등 최신 알고리즘 도입과 그에 따른 임베디드 컴퓨팅 적용 가능성을 분석하여 통합 함상 자동 착륙 모듈을 개발해 군용 드론에 적용할 수 있도록 확대할 예정이다.

ACKNOWLEDGMENTS

이 논문은 2025년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE) (2025-RISE-01-019-04) 사업의 지원과 2025년도 세종대학교 교내연구비 지원에 의한 논문입니다.

AUTHOR CONTRIBUTIONS

Conceptualization, B.C, J.S, and Y.K; methodology, B.C, W.J, and S.J; simulation, B.C, W.J, and M.B; validation, B.C, W.J, M.B, and J.S; formal analysis, B.C, S.J, and J.S; writing-original draft preparation, B.C and J.S; writing-review and editing, B.C, S.J, and Y.K; supervision, J.S and Y.K; project administration, J.S and Y.K; funding acquisition, J.S.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Amendola, J., Cenkeramaddi, L. R., & Jha, A. 2024, Drone Landing and Reinforcement Learning: State-of-Art, Challenges and Opportunities, IEEE Open Journal of Intelligent Transportation Systems, 5, 520-539. https://doi.org/10.1109/OJITS.2024.3444487

Det Norske Veritas 2011, DNV-RP-H103: Modelling and analysis of marine operations, Det Norske Veritas Technical Report, Høvik, Norway, pp.1-200.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. 2018, An introduction to deep reinforcement learning, Found. Trends Mach. Learn., 11, 219-354. https://doi.org/10.1561/2200000071

Kim, J. S. & Shin, S. H. 2019, A study on Shape of Ocean Wave Spectrum, In Proceedings of the Korean Institute of Navigation and Port Research Conference, Jeju, Republic of Korea, 15-17 May 2019, pp.51-52.

Lee, Y. U. 2020, A Study on the Effective Military Use of Drones, J. Converg. Secur., 20, 61-70. https://doi.org/10.33778/kcsa.2020.20.4.061

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., & Erez, T., et al. 2016, Continuous control with deep reinforcement learning, ICLR 2016, Caribe Hilton, San Juan, Puerto Rico, 2-4 May 2016, pp.1-14. https://doi.org/10.48550/arXiv.1509.02971

Matheron, G., Perrin, N., & Sigaud, O. 2020, Understanding failures of deterministic actor-critic with continuous action spaces and sparse rewards, in Artificial Neural Networks and Machine Learning – ICANN 2020 (Cham: Springer International Publishing), pp.308-320. https://doi.org/10.1007/978-3-030-61616-8_25

McTaggart, K. 2012, ShipMo3D Version 3.0 User Manual for Creating Ship Models, Technical Memorandum (Dartmouth: Defence R&D Canada - Atlantic).

Palmer, K. P. C. & Irani, R. A. 2026, Neural Networks for high accuracy short term ship motion predictions with applications to autonomous UAVs, Aerospace Science and Technology, 168, 110964. https://doi.org/10.1016/j.ast.2025.110964

Rodriguez-Ramos, A., Sampedro, C., Bavle, H., de la Puente, P., & Campoy, P. 2019, A Deep Reinforcement Learning Strategy for UAV Autonomous Landing on a Moving Platform, Journal of Intelligent & Robotic Systems, 93, 351-366. https://doi.org/10.1007/s10846-018-0891-8

Subamanian, S. P. V., Subramanian, S. M., Muthiah, P., & Shajahan, J. M. A. 2023, Autonomous Drone Landing on a Moving Naval Base using Vision-Based Robot Control, In Proceedings of 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Canary Islands, Spain, 19-21 July 2023, pp.1-8. https://doi.org/10.1109/ICECCME57830.2023.10252272

Talha, M., Asghar, F., Rohan, A., Rabah, M., & Kim, S. H. 2019, Fuzzy Logic-Based Robust and Autonomous Safe Landing for UAV Quadcopter, Arabian Journal for Science and Engineering, 44, 2627-2639. https://doi.org/10.1007/s13369-018-3330-z

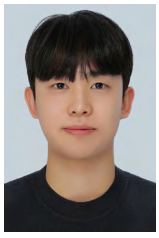
Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J. et al. 2024, Deep Reinforcement Learning: A Survey, IEEE Transactions on Neural Networks and Learning Systems, 35, 5064-5078. https://doi.org/10.1109/TNNLS.2022.3207346

Wu, L., Wang, C., Zhang, P., & Wei, C. 2022, Deep

Reinforcement Learning with Corrective Feedback for Autonomous UAV Landing on a Mobile Platform, Drones, 6, 238. https://doi.org/10.3390/drones6090238

Xie, J., Peng, X., Wang, H., Niu, W., & Zheng, X. 2020, UAV Autonomous Tracking and Landing Based on Deep Reinforcement Learning Strategy, Sensors, 20, 5630. https://doi.org/10.3390/s20195630

Bokyung Choi is M.S Student at Department of Artificial Intelligence and Robotics, Convergence Engineering for Intelligent Drone, Artificial Intelligence and Robotics Institute (AIRI) at Sejong University. She received B.S. degree in Intelligent Mechatronic Engineering from the same university. Her research interests include Inertial Navigation System, Artificial Intelligence for Autonomous Control and Sim-to-Real.



Woo Joo Jung is M.S student at Department of Artificial Intelligence and Robotics and Convergence Engineering for Intelligent Drone, Artificial Intelligence and Robotics Institute (AIRI) at Sejong University. He received B.S. degree in Intelligent Mechatronics Engineering from the same university. His research interests include Inertial Navigation System and Sensor Fusion for Integrated Navigation System.

Min Su Byeon is M.S Student at Department of Artificial Intelligence and Robotics, Convergence Engineering for Intelligent Drone, Artificial Intelligence and Robotics Institute (AIRI) at Sejong University. He received B.S. degree in Aerospace Engineering from the same university. His research interests include Artificial Intelligence for Navigation, Inertial Navigation System and Kalman Filtering, High End Maritime Navigation.

So Yeong Jung is M.S Student at Department of Artificial Intelligence and Robotics, Convergence Engineering for Intelligent Drone, Artificial Intelligence and Robotics Institute (AIRI) at Sejong University. She received B.S. degree in Aerospace Engineering and Intelligent Mechatronic Engineering from the same university. Her research interests include Inertial Navigation System and Kalman Filtering, Dynamic Model Based Navigation, Sensor Fusion and AI-based Navigation System.

Jin Woo Song received the B.S. and M.S. degree in control and instrumentation engineering and the Ph.D. degree in electrical, electronic, and computer engineering from Seoul National University, Seoul, Republic of Korea, in 1995, 1997, and 2002 respectively. He is currently an Associate Professor with the department of artificial intelligence and robotics, and the department of convergence major for intelligent drones at Sejong University, Seoul, Republic of Korea. He is currently a member of Artificial Intelligence and Robotics Institute. His research interests include GNSS/INS, robust and optimal control, and MEMS sensors.

Yong Hun Kim received the B.S. degree in robotics engineering from Hoseo University, Asan-si, Republic of Korea, in 2018. He received the M.S. degree in software convergence and the Ph.D. degree in intelligent mechatronics engineering with a convergence major in intelligent drones from Sejong University, Seoul, Republic of Korea, in 2020 and 2024, respectively. He is currently an Assistant Professor in the department of artificial intelligence and robotics at Sejong University. He is currently a member of Artificial Intelligence and Robotics Institute. His research interests include nonlinear filtering, indoor navigation, personal navigation systems, and multi-sensor integration.