대한조선학회논문집 Research Paper
Journal of the Society of Naval Architects of Korea
eISSN:2287-7355, Vol. 63, No. 3, pp. 149-161, June 2026
https://doi.org/10.3744/SNAK.2026.63.3.149

다중 에이전트 강화학습을 적용한 유·무인 복합 수중 운동체의 편대 전환 프레임워크 초기 개발

김승태[1]·유영준[1,2,†]
울산대학교 일반대학원 조선및해양공학과[1]
울산대학교 공과대학 조선해양공학부[2]

Preliminary Development on Framework to Execute Formation Transformation for Manned-unmanned Teaming Underwater Vehicles by Implementing Multi-agent Reinforcement Learning

Seungtae Kim[1]·Youngjun You[1,2,†]
Graduate School of Naval Architecture and Ocean Engineering, University of Ulsan[1]
School of Naval Architecture and Ocean Engineering, University of Ulsan[2]

Recently, the paradigm of defense technology has been shifting from manned weapon systems to distributed systems, including Manned-Unmanned Teaming(MUM-T) enabled by artificial intelligence technologies. In maritime industries, MUM-T has been studied for both surface vessels and underwater vehicles. However, limitations in visibility, communication, and operational depth in underwater environments are regarded as critical barriers for implementing MUM-T technologies in underwater vehicles. Furthermore, it is challenging to establish an initial concept for cooperative underwater vehicles based on MUM-T. In this paper, operation technology for MUM-T in cooperative underwater vehicles was developed. In particular, a framework for conducting formation transformation of a fleet was proposed. First, studies on MUM-T formations being developed in various fields, including underwater vehicles, were investigated. Second, multi-agent reinforcement learning was applied to the methodology for MUM-T formation keeping and transformation. Third, the fair Hungarian algorithm was applied to the formation transformation problem to improve the operational duration of the entire fleet. Fourth, the formation transformation results of MUM-T were analyzed and examined using the proposed multi-agent reinforcement learning model.

Keywords : Formation transformation(편대 전환), Manned-Unmanned Teaming(유·무인 복합체계), Underwater vehicles(수중 운동체), Multi-Agent reinforcement learning(다중 에이전트 강화학습)

1. 서 론

최근 인공지능, 무인 로봇 무기의 개발 및 운용이 가속화됨에 따라, 국가 간 안보 패러다임도 빠르게 변화하고 있다(Lee, 2022). 우크라이나-러시아, 이란-미국, 이스라엘 간 전쟁에서는 무인체계가 단순한 보조 전력이 아니라 감시, 정찰, 표적획득, 타격, 기만, 및 소모전 등을 수행하는 핵심 전력으로 자리매김하고 있음을 보여주며, 유·무인 전력 간의 협력이 전장의 효과성과 지속성에 매우 중요함을 인식하게 되었다(NATO, 2025). 따라서, 방위산업 기술의 패러다임이 단일 고성능 유인 플랫폼 중심에서, 자율성, 연결성, 분산성에 기반한 유·무인 복합 전투 체계 중심으로 전환될 것으로 예상된다. 즉, 유인 체계의 생존성과 판단 능력, 그리고 무인체계의 저위험, 저비용, 대량 운용 특성이 결합한 유·무인 복합체계(Manned-Unmanned Teaming, MUM-T)는 차세대 전장 운용 개념의 핵심으로 주목받고 있다(U.S. Army, 2010; U.S. Department of Defense, 2011). NATO에서도 인공지능과 자율 체계가 군 전반의 작전 수행 방식을 변화시킬 핵심 기술임을 공표하였으며, 대한민국 국방부에서도 인공지능 기반

Received : 20 April 2026 | Revised : 4 May 2026 | Accepted : 11 May 2026
† Corresponding author : Youngjun You, youngjunyou@ulsan.ac.kr

유·무인 복합 전투 체계 발전 방안을 제시한 바 있다(NATO, 2021; 대한민국 국방부, 2023).

이에 따라 유·무인 복합체계에 관한 연구는 공중, 지상, 해양(수상, 수중) 방산기술 분야에서 다양하게 수행되고 있다. Xing et al. (2024)는 불확실한 장애물 환경에서 다수의 무인항공기(Unmanned Aerial Vehicle, UAV) 편대의 경로계획을 위해 강화학습 기법을 적용한 바 있다. Mead et al. (2026)은 다단계 학습과 다중 에이전트 강화학습을 결합하여 다수의 무인지상차량(Unmanned Ground Vehicle, UGV) 경로계획과 협력 임무 수행 과정을 무인 지상차량을 이용하여 검증하였다. Zhao et al. (2021)은 다수의 무인 수상 운동체(Unmanned Surface Vehicle, USV)의 편대 유지 및 경로 추종을 위해 강화학습을 적용하였고, 일부 개체가 편대를 이탈했을 때 편대를 재구성함으로써 임무의 지속성을 보인 바 있다.

앞에서 기술한 공중, 지상, 해양(수상)에서의 유·무인 복합체계 연구·개발과 다르게, 해양(수중) 환경은 제한적인 가시성, 불완전한 통신, 운용 수심에 따른 제약 등으로 수중 운동체의 유·무인 복합체계 운용 개념을 정립하는 데 상당한 어려움이 따른다. Liu et al. (2023)은 불완전한 수중 통신 환경에서 간헐적으로 획득되는 위치 정보를 활용하여, 무인 수중 운동체의 경로 추종을 수행하기 위하여 강화학습 기반 제어 기법을 제안하였다. Cao et al. (2025)는 복잡한 해양(수중) 환경에서 다수의 무인 수중 운동체(Unmanned Underwater Vehicle, UUV)의 협력 임무 중 충돌 회피를 위하여 다중 에이전트 강화학습을 적용한 바 있다. 기존 연구에서는 유인 에이전트와 무인 에이전트 사이의 우선순위를 부여하거나, 각각의 에이전트가 가지고 있는 전략 자산의 중요도에 따른 우선순위를 고려하고 있지 않기 때문에, 유·무인 복합체계의 운용 개념을 구현하는데 한계가 있다고 판단하였다. 따라서, 임무 및 운용 환경 변화에 대응할 수 있도록 다층적인 편대를 고려한 군집 수중 운동체 편대 전환 또는 유지에 관한 프레임워크 개발 필요성을 인지하였다.

본 연구에서는 다중 에이전트 강화학습을 적용하여, 유·무인 복합 수중 운동체의 편대 전환 및 유지에 관한 프레임워크를 개발하고, 적용 가능성을 검토하고자 했다. 먼저, 수중 운동체를 포함한 다양한 분야에서 개발되고 있는 유·무인 복합체계 편대 관련 연구를 분석 및 검토하였다. 둘째, 유·무인 수중 운동체의 편대 전환 및 유지 방법론에 있어, 다중 에이전트 강화학습을 적용하였다. 각각의 에이전트를 점 질량으로 가정한 후, 환경, 상태, 행동 및 보상함수를 포함한 강화학습 변수를 설계하였다. 셋째, 군집 수중 운동체의 편대 전환 시 이동 거리의 편차를 최소화할 수 있도록, 공평한 헝가리안 알고리즘을 적용, 검증하였다. 마지막으로, 제안된 다중 에이전트 강화학습 모델을 이용하여 유·무인 복합 수중 운동체의 편대 전환 결과를 분석, 검토하였다.

본 논문의 구성은 다음과 같다. 2장에서 수중 운동체의 수학모형, 적용된 다중 에이전트 강화학습 기법을 설명하였다. 3장에서는 다중 에이전트 강화학습에 필요한 모델을 세부적으로 기술하고자 했다. 특히, 환경을 병렬화한 다중 에이전트 강화학습 기법과 설계된 입력 변수, 출력 변수 및 보상함수를 설명하고 있다.

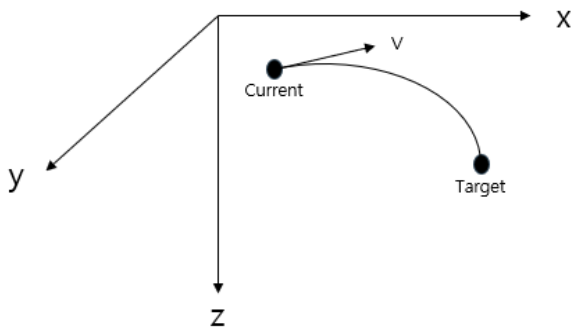

Fig. 1 Coordinate system of point mass models

4장에서는 제안된 다중 에이전트 강화학습을 적용하여, 편대 전환 시뮬레이션을 수행한 후, 시뮬레이션 결과의 타당성을 분석, 검토하였다. 5장에서는 본 연구에서 검토한 사항을 요약, 정리하였다.

2. 대상 개체 및 수학 모형

2.1 대상 개체 및 축계 정의

본 연구에서는 각각의 수중 운동체를 Fig. 1에서 보는 것처럼 점질량(point mass)으로 가정하였다. 통상, 수중 운동체의 유체동역학 분야 연구에서는 강체(rigid body)를 가정한 후, 6 자유도 운동을 풀이하는 것이 일반적이다. 본 저자는 다수의 유무인 복합 수중 운동체를 대상으로 편대 전환을 수행하기 위한 다중 에이전트 강화학습 기반 프레임워크의 초기 개발을 완료한 후, 6자유도 운동까지 고려하는 단계적인 연구·개발 계획을 수립하였다. 여기서는 유·무인 복합 수중 운동체의 편대 전환 프레임워크의 타당성 및 적용 가능성 검토를 위한 1단계 개발로써, 각각의 에이전트를 점질량으로 단순화하였다.

2.2 공평한 헝가리안 알고리즘 및 검증

다수의 유·무인 복합 수중 운동체의 운용 개념을 구상하는 데 있어, 잠수함과 같은 대형 유인 에이전트뿐만 아니라 배터리 기반 전기추진 시스템을 탑재한 중·소형 에이전트도 고려할 필요가 있었다. 특히, 수중 환경은 통신 제약, 제한된 에너지, 임무 중 재충전의 어려움 등이 있어, 특정 에이전트에 과도한 이동 거리가 요구될 경우 전체 편대 전환 및 유지에 관한 협력 임무 수행의 안정성과 지속성이 저하될 위험성이 크다(Islam et al., 2022). 일반적으로 헝가리안 알고리즘은 군집 운동체의 편대 전환 및 유지 시 전체 이동 거리의 합계 최소화를 목표하기 때문에, 일부 에이전트에 과도한 이동 거리가 요구될 수 있다. Moon (2022)는 에이전트 간 이동 거리를 균등하게 가져가기 위하여, 공평한 헝가리안 알고리즘을 제안한 바 있다. 본 연구에서는 공평한 헝가리안 알고리즘을 이용하여, 유·무인 복합 수중 운동체의 편대 전환 시 개체 간 이동 거리의 편차를 최소화하고자 했다.

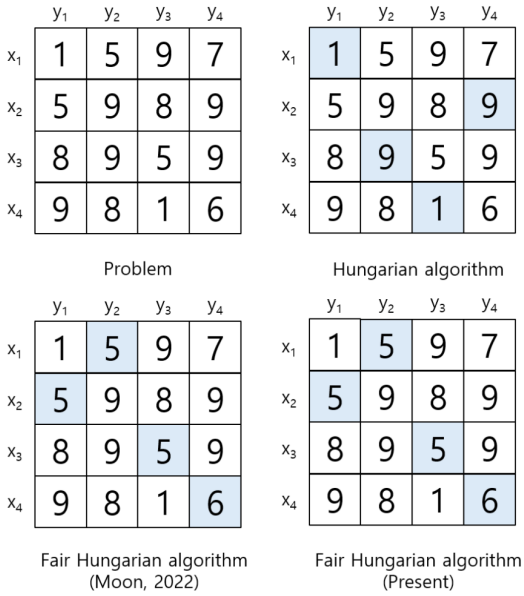

Fig. 2 Comparisons among obtained results from Hungarian algorithm, Fair Hungarian algorithm (Moon, 2022) and Fair Hungarian algorithm (present study)

일반적으로 헝가리안 알고리즘은 이분 그래프 기반 할당 문제로 정식화되며, 식 (1)에서 보는 것처럼 시작 위치와 목표 위치 사이의 간선 가중치를 이동 거리로 정의한다. 이후 식 (2)와 같이 할당된 간선 가중치의 총합이 최소가 되도록 목적함수를 구성한다. 여기서, e는 현재 위치(x)와 목표 위치(y)를 연결하는 간선이고, w는 각 간선의 가중치, p는 에이전트의 3차원 위치를 나타낸다. i는 현재 위치의 에이전트 번호, j는 목표 위치의 에이전트 번호를 가리킨다. 반면 공평한 헝가리안 알고리즘은 식 (3)과 같이 적용 가능한 범위 내에서 최대 이동 거리를 최소화하는 방식으로 목적함수를 재구성한다. 즉, 전체 이동 거리의 증가를 최소화하면서도 특정 에이전트의 과도한 이동이 집중되지 않도록 설계한다. 또한 식 (4)와 같이 간선 가중치의 최댓값을 차례대로 제거하면서 홀 정리를 이용하여 집합 X의 원소가 집합 Y의 원소와 일대일 대응하는 것을 확인할 수 있다. 여기서 X는 현재 위치의 전체 집합, Y는 목표 위치의 전체 집합이다. S는 X의 부분 집합이고, N(S)은 S에 속한 원소들이 연결될 수 있는 목표 위치의 이웃 집합을 가리킨다.

Moon (2022)이 제시한 할당 문제를 대상으로 기존 헝가리안 알고리즘을 적용한 결과, 공평한 헝가리안 알고리즘의 결과, 본 연구에서 공평한 헝가리안 알고리즘을 적용한 결과를 각각 비교함으로써, 본 연구에서 공평한 헝가리안 알고리즘의 타당성을 검증하고자 했다. Fig. 2에서 에이전트의 현재 위치와 목표 위치 사이의 이동 거리를 각각의 칸에 표시하였다. 각각의 에이전트가 최종적으로 선택한 조건은 파란색 음영으로 표시하였다. 공평한 헝가리안 알고리즘을 적용한 경우, 총 이동 거리는 20m로 헝가리안 알고리즘 대비 약 5% 증가하였으나, 이동 거리의 표준편차는 0.5로 약 89.1% 감소하였다. 이동 거리의 합계는 미미하게 증가하지만, 개별 개체 간 이동 거리 편차는 뚜렷하게 감소하였

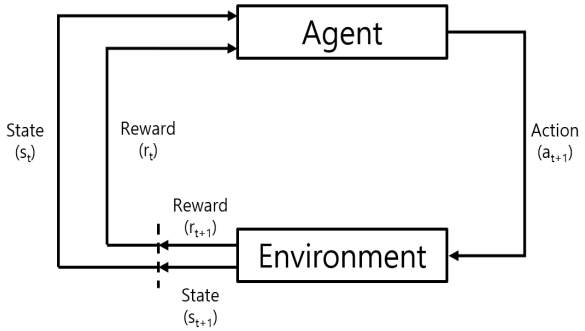

Fig. 3 S/W architecture of single-agent reinforcement learning (Sutton and Barto, 2018)

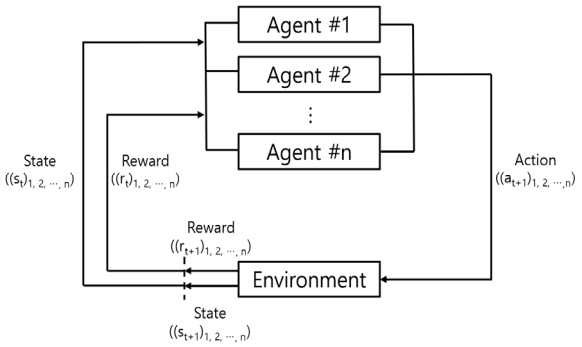

Fig. 4 S/W architecture of multi-agent reinforcement learning (Sutton and Barto, 2018)

음을 가리킨다. 또한, 적용된 공평한 헝가리안 알고리즘의 결과를 Moon (2022)의 결과와 비교했을 때, 같은 결과를 얻었음을 확인할 수 있다. 이를 통해 본 연구에서 적용된 공평한 헝가리안 알고리즘을 검증할 수 있었다.

$$w(e_{x,y}) = \sqrt{\{(p_i)_x - (p_j)_x\}^2 + \{(p_i)_y - (p_j)_y\}^2 + \{(p_i)_z - (p_j)_z\}^2} \quad (1)$$

$$\min.\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} w(e_{xy})\right\} \quad (2)$$

$$\min.\left[\max_{x\in X,\, y\in Y}.\{w(e_{xy})\}\right] \quad (3)$$

$$|S| \le |N(S)| \text{ for } S \subseteq X \quad (4)$$

2.3 다중 에이전트 강화학습

강화학습은 에이전트가 환경과 상호작용을 하면서 현재 상태에서의 행동을 선택하고, 그 결과로 획득한 보상을 누적하여 보상값을 최대화하도록 정책을 학습하는 기법이다(Sutton and Barto, 2018). 일반적으로, 강화학습은 단일 에이전트 강화학습과 다중 에이전트 강화학습으로 구분된다. Fig. 3과 같이 단일 에

이전트 강화학습은 하나의 에이전트가 단일 환경 내에서 상태를 관측한 후 행동을 선택하며, 그에 따른 보상을 통해 정책을 학습한다(Shin et al., 2020). 반면 다중 에이전트 강화학습은 Fig. 4와 같이 둘 이상의 에이전트가 동일한 환경 내에서 동시에 상호작용을 하며 각자의 정책을 학습한다(Zhang et al., 2026).

다중 에이전트 강화학습 알고리즘은 일반적으로 가치 기반(value-based), 정책 기반(policy-based), 행위자-비평자 기반(actor-critic-based)으로 구분된다 (Ning and Xie, 2024). 이 중 행위자-비평자 기반 방법은 정책을 학습하는 행위자(actor)와 가치 함수를 추정하는 비평자(critic)를 함께 학습함으로써, 가치 기반과 정책 기반의 장점을 결합한 것이 특징이다. 특히, 복잡한 연속 제어 문제에서 행위자-비평자 기반 알고리즘이 널리 활용되고 있다(Zhang and Yu, 2020). 본 연구에서도 연속적인 행동에 적합한 행위자-비평자 기반 강화학습을 적용하였다.

행위자-비평자 기반 강화학습은 정책 데이터를 활용하는 방식에 따라 on-policy와 off-policy로 구분된다. On-policy 방법은 현재 정책으로부터 수집한 데이터를 이용하여 정책을 갱신하는 반면, off-policy 방법은 과거 또는 다른 정책으로 수집한 데이터도 재사용하여 학습할 수 있다(Gorji and Granmo, 2023). 그러나 off-policy 방법은 현재 갱신하려는 정책과 데이터가 수집된 시점의 정책 사이에 차이가 존재하므로, 정책 불일치로 인한 학습 결과의 불안정성 문제가 발생할 수 있다. 특히, 다중 에이전트를 고려할 경우, 각 에이전트의 정책이 동시에 변화하기 때문에 비정상성 문제와 학습 결과의 불안정성에 관한 우려가 제기되었다. 반면 on-policy 방법은 현재 정책에 기반한 최신 데이터를 사용하므로 학습 안정성이 상대적으로 높고, 다중 에이전트 환경에서 발생하는 비정상성을 해결하는 데 유리하다(Ning and Xie, 2024). 본 연구에서는 on-policy 방법이 갖는 장점이 본 논문에서 목표하고 있는 프레임워크 개발에 필요하다고 판단하였다.

PPO(Proximal Policy Optimization)는 정책 경사 기반의 대표적인 on-policy 강화학습 알고리즘으로, Table 1에서 설명하고 있는 것처럼 환경과의 상호작용을 통해 수집된 표본으로부터 대리 목적함수(surrogate objective)를 최적화하여 정책을 갱신한다 (Nam et al., 2023). 정책 갱신 과정에서 clipping 기법을 도입하여 새로운 정책이 기존 정책으로부터 과도하게 벗어나지 않도록 제한한다. 식 (5)는 현재 정책과 이전 정책 간 차이를 비율로 표시한 것이며, 식 (6)은 PPO에서 이용되는 대리 목적함수를 정의한 것이다. 이러한 clipping 기법은 정책 업데이트 폭을 제한하여 급격한 성능 저하를 방지하고, 학습 수렴성을 향상할 수 있다. 본 연구에서는 다수의 수중 운동체가 동시에 위치를 이동하는 협력적 편대 전환 문제를 풀이하는 데 있어, PPO 알고리즘이 갖는 안정성과 수렴성이 필요하다고 판단하였다.

PPO는 행위자-비평자 구조를 기반으로 하므로, 정책을 갱신하는 행위자와 상태 가치를 추정하는 비평자를 함께 학습한다. 이때 비평자는 가치추정값과 반환값의 차이를 최소화하도록 동작하며, 식 (7)과 같다. 식 (8)은 특정 시점에서의 실제 반환값이 비평자가 예측한 상태 가치보다 얼마나 우수한지를 나타내며, 정책이 어떤 행동을 선택해야 하는지를 결정하는 기준으로 활용된다.

Table 1 Pseudo-code of multi-agent proximal policy optimization algorithm

Algorithm 1 MAPPO

Initialize decentralized actor parameters θ
Initialize centralized critic parameters ϕ
for $episode = 1, 2, \cdots, episode_{\max}$ do
  Initialize rollout buffer $D \leftarrow \varnothing$
  Set old actor parameter $\theta_{old} \leftarrow \theta$
  for $t = 1, 2, \cdots, t_{\max}$ do
    for $i = 1, 2, \cdots, N$ do
      Receive local observation $o_{i,t}$
      Sample action $a_{i,t} \sim \pi_{\theta_{old,i}}(a_{i,t}|o_{i,t})$
    end for
    Execute joint action $a_t = (a_t)_1, \cdots, (a_t)_N$
    Receive reward r_t, global state s_{t+1}, and next local observations o_{t+1}
    Store $(s_t, o_t, a_t, r_t, s_{t+1}, o_{t+1})$ in D
  end for
  Compute returns $\hat{R}_t$
  Compute advantages $\hat{A}_t$
  for $k = 1, 2, \cdots, k_{\max}$ do
    Sample mini-batch $B \subset D$
    for $i = 1, 2, \cdots, N$ do
      Optimize actor objective $L_i^{CLIP}(\theta_i)$
    end for
    Optimize critic objective $L_i^{VALUE}(\phi)$
  end for
end for

식 (9)는 정책이 특정 행동에 치우치지 않도록, 탐색을 수행하도록 한다. 최종적으로 PPO의 목적함수는 정책 손실, 가치 함수의 손실, 탐색 성능 유지를 위한 엔트로피를 함께 고려하며, 식 (10)과 같이 얻을 수 있다.

본 연구에서 설계한 MAPPO의 구조는 Fig. 5와 같다. MAPPO는 중앙집중 학습-분산 실행(Centralized Training and Decentralized Execution, CTDE) 구조를 적용한다(Tan et al., 2026). 학습 단계에서는 비평자가 다수 에이전트의 상태 및 행동 정보를 종합적으로 활용하여 가치 추정을 수행하고, 실행 단계에서는 각 에이전트가 자신의 관측 정보만을 이용하여 독립적으로 행동을 결정한다. 이를 통해, 학습 시에는 다중 에이전트 간 협력을 충분히 반영할 수 있고, 운용 시에는 통신 및 관측 제약이 존재하는 분산 환경에도 대응할 수 있다는 장점이 있다. 따라서, 다수의 에이전트 간 협력을 고려하면서도, 개별 에이전트의 독립적 판단이 요구된다는 점에서 MAPPO를 이용하는 것이 주어진 유·무인 복합 수중 운동체 편대 전환 문제에 필요하다고 판단하였다.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{5}$$

$$L^{CLIP}(\theta) = \hat{E}_t[\min.(r_t(\theta)\hat{A}_t,\ clip(r_t(\theta),\ 1-\epsilon,\ 1+\epsilon)\hat{A}_t] \tag{6}$$

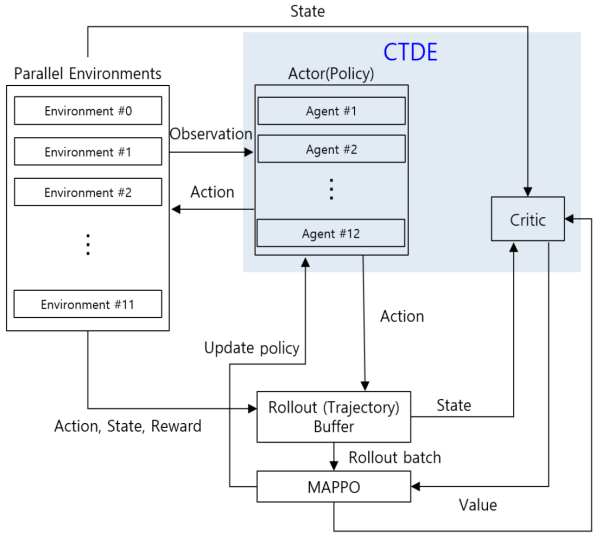

Fig. 5 S/W architecture of multi-agent proximal policy optimization

$$L^{VALUE}(\phi)=\hat{E}_t[(V_\phi(s_t)-\hat{R}_t)^2] \tag{7}$$

$$\hat{A}_t=\hat{R}_t-V_\phi(s_t) \tag{8}$$

$$L^{ENTROPY}(\theta)=\hat{E}_t[-\sum_a \pi_\theta(a_t|s_t)\log\pi_\theta(a_t|s_t)] \tag{9}$$

$$L(\theta,\phi)=L^{CLIP}(\theta)-c_1L^{VALUE}(\phi)+c_2L^{ENTROPY}(\theta) \tag{10}$$

3. 다중 에이전트 강화학습 모델링

3.1 강화학습 환경 병렬화

다중 에이전트 강화학습 기법을 이용하여 유·무인 복합 수중 운동체의 편대 전환을 풀이하기 위하여, 먼저 편대 형상을 정의하였다. 편대 형상은 수중 운동체의 공간적 배치 특성을 고려하여 1차원 편대(직선), 2차원 수평면 편대(타원형), 2차원 수직면 편대(타원형), 3차원 편대(타원체) 총 네 가지로 정의하였다. 1차원 편대는 기지의 출입 또는 항만 내 이동, 2차원 수평면, 수직면 편대는 해저 지형 또는 수심 제한에 따른 이동, 3차원 편대는 협력 임무 수행을 위한 정상 운항을 각각 가정한 것이다. 모든 편대 형상에서 1개의 유인 에이전트는 편대의 중심에 위치함으로써, 유인 에이전트를 중심으로 편대 전환 과정을 모사하고자 했다. 그 외, 무인 에이전트는 기하학적 편대 형상을 따라, 내부의 다각형 또는 다면체의 꼭짓점에 배치되었다고 가정하였다.

모든 편대는 유인 에이전트를 중심으로 2개의 층(내부, 외부)으로 구성된다. 유·무인 복합 군집 수중 운동체를 구성하는 데 있어, 유인 에이전트를 포함하여 전략 자산의 가치에 따라 중요도가 상대적으로 높은 에이전트를 내부에 배치하고, 그 외 무인 에이전트를 외부에 배치하였다. 각 에이전트의 중요도는 유인 체계 여부, 미사일 또는 탑재된 전략 자산의 경제적 가치 등에 따라 달라진다. 여기서, 유인 수중 운동체를 제외한 4개의 1차 무인 에이전트(내부), 8개의 2차 무인 에이전트(외부)로 총 12개의 에이전트로 가정하였다. 각 에이전트는 강화학습 환경 내에서 독립적으로 행동을 하면서도, 편대 전환이라는 목표를 달성하도록 행동한다. 특히, 편대 전환 과정에서도 에이전트의 중요도는 바뀌지 않기 때문에, 내부, 외부 층을 유지하도록 했다.

Table 2 Definition of underwater vehicles' formations

Formation	Specifications					
	Longitudinal length [m]		Transverse length [m]		Vertical length [m]	
	Inner layer	Outer layer	Inner layer	Outer layer	Inner layer	Outer layer
1D (line)	40	120	–	–	–	–
2D horizontal (ellipse)	60	100	42	70	–	–
2D vertical (ellipse)	60	100	–	–	30	50
3D (ellipsoid)	60	100	42	70	30	50

Table 2에 제시된 편대의 기하학적 형상 정보는 점질량으로 단순화된 에이전트를 기준으로 가정하였다. 1차원 편대(직선)는 제한된 공간에서 종 방향으로 이동하는 상황을 고려하여 12개의 에이전트와 1개의 리더가 종 방향으로 충분한 안전거리를 확보할 수 있도록 10m 간격으로 배치하였다. 2차원 편대(타원형)는 해저 지형 또는 수심 제한 등으로 인해 수평면 또는 수직면에서 편대 이동이 필요한 상황을 고려하였다. 따라서 1차원 편대(직선)보다 넓은 작전 공간에서 편대 유지 및 임무 수행을 위해, 상대적으로 큰 종 방향 길이를 갖도록 설정하였다. 이에 따라 내부 층과 외부 층 종 방향 길이는 각각 60m, 100m로 가정하였다. 또한 수중 운동체는 일반적으로 횡 방향 및 상하 방향 길이가 종 방향 길이보다 짧은 형상적 특성을 가지므로, 2차원 수평면(타원형) 편대의 횡 방향 길이는 종방향 길이의 70%로 설정하였다. 2차원 수직면(타원형) 편대의 경우는 수중 운동체의 심도 변화에 대한 민감성과 수직 방향 운용 제약을 고려하여 상하 방향 길이를 종 방향 길이의 50%로 설정하였다. 3차원 편대(타원체)도 2차원 편대(타원형)과 동일하게 장축은 내부 층 60m, 외부 층 100m로 설정하고 단축은 장축의 70%, 높이는 장축의 50%로 설정하였다. 이를 통해 3차원 공간에서 수중 운동체의 종 방향으로 긴 형상적 특성과 수직 방향 운용 제약을 동시에 반영하고자 하였다. 본 연구에서는 다중 에이전트 강화학습을 적용한 프레임워크 구축 및 타당성 검토까지 목표로 했기 때문에, 정의된 변수를 정식화하는 데 초점을 두었다. 향후 후속 연구 단계에서, 실제 수중 운동체 운용 개념을 고려하여, 편대 형상과 관련된 정량화 연구를 다루어야 할 것으로 판단된다.

초기 편대 형상에서부터 다른 목표 편대 형상으로 전환하는 과

Table 3 Definition of parallelized multi-agent reinforcement learning environments

Items	Specifications
Environment #0	1D(line) to 2D horizontal(ellipse)
Environment #1	1D(line) to 2D vertical(ellipse)
Environment #2	1D(line) to 3D(ellipsoid)
Environment #3	2D horizontal(ellipse) to 2D vertical(ellipse)
Environment #4	2D horizontal(ellipse) to 3D(ellipsoid)
Environment #5	2D horizontal(ellipse) to 1D(line)
Environment #6	2D vertical(ellipse) to 3D(ellipsoid)
Environment #7	2D vertical(ellipse) to 1D(line)
Environment #8	2D vertical(ellipse) to 2D horizontal(ellipse)
Environment #9	3D(ellipsoid) to 1D(line)
Environment #10	3D(ellipsoid) to 2D horizontal(ellipse)
Environment #11	3D(ellipsoid) to 2D vertical(ellipse)

정을 하나의 강화학습 환경으로 정의하였다. 예를 들어, 1차원 편대(직선)에서부터 3차원 편대(타원체)로 전환은 하나의 독립된 환경에 해당한다. 모든 형상 간 편대 전환을 고려하면, Table 3에서 보는 것처럼 모두 12개의 경우의 수가 존재한다. 하나의 편대 전환 시나리오를 하나의 환경 단위로 정의한 것은 편대 전환 시 초기 형상과 목표 형상에 따라 서로 다른 차원을 가질 수 있기 때문이다. 예를 들어, 1차원 편대(직선)에서부터 2차원 수평면 또는 2차원 수직면 편대로의 전환, 3차원 편대(타원체)로의 전환에서 차원이 달라질 수 있다. 따라서, 하나의 단일 환경으로 고려하는 것보다, 편대 전환을 각각의 환경 단위로 정의하는 것이 편대 전환 특성을 분석, 검토하기 쉽다고 판단하였다.

강화학습에서 환경 병렬화는 다수의 환경을 동시에 실행하여 경험 표본을 병렬적으로 수집하는 방식을 가리키며, 단일 환경을 차례대로 반복하는 것보다 학습 효율을 크게 향상할 수 있다(Hou et al., 2022). 첫째, 여러 환경에서 동시에 trajectory(행동, 상태, 보상함수)를 생성할 수 있으므로 단위 시간당 수집되는 경험의 양이 증가하여 학습 속도를 높일 수 있다. 둘째, 서로 다른 편대 전환 환경으로부터 다양한 상태와 보상 정보를 동시에 확보할 수 있어, 학습 데이터의 다양성이 증가하고 정책의 일반화 성능을 향상할 수 있다. 셋째, 특정 환경에 과도하게 편향된 경험 축적을 방지함으로써, 정책이 일부 형상 전환에만 특화되는 현상을 해결할 수 있다. 넷째, 환경 병렬화는 on-policy 기반인 MAPPO에서 rollout 수집 과정을 효율화하여, 정책 업데이트에 필요한 표본을 더 안정적으로 확보할 수 있다.

3.2 강화학습 변수 및 보상함수 정식화

강화학습을 적용하는 데 있어, Table 4에서 보는 것처럼 행동, 관측, 상태에 관한 변수를 설계할 필요가 있었다. 여기서, 행동은 각 에이전트가 환경 내에서 선택할 수 있는 제어 입력의 범위를

Table 4 Action, observation, and state

Items	Specifications
Action	$u[kts]\in[0,4]$ $v[kts]\in[-2,2]$ $w[kts]\in[-2,2]$
Observation	$o_i=\left\{d_i^{Goal}, d_i^L, d_i^j, p_L, p_i^j, V_i^j, z-z_{\min}, z_{\max}-z\right\}$
State	$s=\left\{R_F, p_L, p_i, V_i, d_i^j,\right\}$

의미한다. 관측은 개별 에이전트가 의사 결정을 위해 직접 획득할 수 있는 로컬 정보를 가리킨다. 상태는 환경 전체를 대표하는 정보로, 특히 MAPPO의 CTDE 구조에서는 비평자가 학습 단계에서 활용하는 전역 정보를 가리킨다(Tian et al., 2024). 따라서 관측은 로컬 정보에 기반하여 분산 실행을 수행하는 데 이용하며, 상태는 다수 에이전트 간 상호작용과 편대 전환 과정을 반영하여 중앙집중 학습을 수행하는 데 이용된다.

먼저, 각 에이전트의 행동은 수중 운동체의 점질량 기반 이동을 고려하여 x, y, z 방향 속도 u ,v, w를 제어 입력으로 정의하였다. 종 방향 속도(u)는 수중 운동체의 추진력을 고려하여 가장 큰 범위로 설정하였다. 여기서, 종 방향 속도의 최솟값을 0으로 설정한 것은 편대 전환 과정에서 전진, 후진 모드 전환에 따른 물리적 제약과 소요 시간의 증가를 고려하기 위해서였다. 횡 방향 속도(v)와 상하 방향 속도(w)는 종 방향 속도에 비해 작은 값을 갖는 6 자유도 동특성을 고려하였다.

각 에이전트의 관측은 목표 위치까지의 거리, 리더(유인 에이전트)의 위치, 리더까지의 거리, 가장 가까운 3개 에이전트와의 상대 위치, 상대 속도 및 거리, 그리고 최소 및 최대 운항 심도를 포함하도록 설계하였다. 먼저 목표 위치까지의 거리, 리더의 위치를 관측함으로써 리더를 따라 편대를 전환할 수 있도록 가정하였다. 각각의 에이전트, 또는 리더와 에이전트 사이의 충돌을 방지하기 위해, 편대 전환 중 가장 가까운 3개의 에이전트를 대상으로 상대 위치, 상대 속도, 상대 거리를 관측하였다. 마지막으로, 허용 가능한 운항 심도 이탈 방지를 위하여, 편대 전환 중 현재 심도와 최소 운항 심도의 차이, 최대 운항 심도와 현재 심도의 차이를 관측하였다. 특히 관측을 전역 정보가 아닌 로컬 정보로 제한한 것은 MAPPO의 분산 실행을 위한 것이며, 수중 환경에서 각각의 에이전트가 전체 편대에 속한 다른 에이전트의 모든 정보를 실시간으로 공유받기 어려움을 고려한 것이다.

상태는 학습 단계에서 중앙집중 비평자가 사용하는 전역 정보로써, 편대 전환 진행률, 리더 위치, 모든 에이전트의 현재 위치, 현재 속도, 그리고 목표 위치로 설계하였다. 먼저 편대 전환 진행률은 임의의 시점까지 이동한 거리를 목표 전환 거리로 나눈 값으로 정의하며, 각 에이전트가 목표 편대 전환을 진행한 비율을 가리킨다. 이 값을 고려함으로써, 비평자는 동일한 위치 및 속도 조건이라도 편대 전환의 초기, 중간 또는 최종 시점에 따라 다른 가치 평가를 수행할 수 있다. 리더 위치를 통해 유인 에이전트를 중심으로 임무가 부여되는, 대상 에이전트 중 가장 중요도가 높은 유·무인 복합체계의 운용 개념을 반영하고자 했다. 모든 에이전트의 현재 위치, 현재 속도 및 목표 위치는 다중 에이전트 환경

Table 5 Reward functions

Item	Reward functions
#1	$r_1 = d_i(t-1) - d_i(t)$ $where, d_i(t) = \parallel \{p_i(t) - p_L(t)\} - p_{Goal}(t) \parallel$
#2	$-0.01 \parallel v_i(t) \parallel$
#3	$\begin{cases} -50, (\min.(d_i^j(t)) < R_{AA} \text{ or } \min.(d_i^L(t)) < R_{LA} \\ 0, \quad otherwise \end{cases}$
#4	$\begin{cases} -1, z_i(t) < z_{\min} \text{ or } z_i(t) > z_{\max} \\ 0, \quad otherwise \end{cases}$
#5	$\begin{cases} 1, \text{ if } \max.(d_i^{Goal}(t)) < 1 \\ 0, otherwise \end{cases}$

에서 발생하는 비정상성을 완화하고, 편대의 전환 과정을 평가하기 위해 포함하였다. 편대 전환 문제는 개별 에이전트가 단순히 목표 위치로 이동하는 문제가 아니라, 다수의 에이전트가 상호작용하면서 목표 편대 형상을 형성하는 협력 문제이다. 따라서 중앙집중 비평자는 모든 에이전트의 위치와 속도를 바탕으로 초기 편대 형상에서 목표 편대 형상으로의 전환하는 과정, 에이전트 간 상대적 배치 등이 갖는 안정성을 평가할 수 있다. 목표 위치는 현재 편대 형상과 목표 편대 형상 사이의 차이를 평가하는 데 이용하였다. 이를 통해 비평자는 각 에이전트의 현재 위치가 목표 위치에 얼마나 근접했는지뿐만 아니라, 전체 에이전트가 목표 편대 형상으로 수렴하고 있는지를 종합적으로 판단할 수 있다. 이를 통해 개별 행위자가 더욱 협력적인 방향으로 정책을 학습할 수 있도록 한다.

보상함수는 에이전트가 수행한 행동의 결과를 정량적으로 평가, 제공하는 함수이며, 에이전트는 보상함수로 정의된 누적 보상의 기댓값을 최대화하도록 정책을 학습한다(Sutton and Barto, 2018). 유·무인 복합 수중 운동체의 편대 전환 문제에서는 각 에이전트가 목표 위치로 이동하는 것뿐만 아니라, 전환 과정에서 편대를 안정적으로 유지하고, 리더 및 인접 에이전트와의 충돌을 방지하며, 수중 운동체에 요구되는 심도 내에서 운용되어야만 한다. 또한 개별 에이전트의 목표 위치에 도달한 것만으로 편대 전환이 성공적으로 완료되었다고 볼 수 없기 때문에, 모든 에이전트가 목표 위치에 도달했는지를 함께 고려할 필요가 있었다. 본 연구에서는 목표 위치 추종, 급격한 가감속으로 인한 과도한 운동 방지, 충돌 회피, 운항 심도 준수, 그리고 편대 전환 성공 여부를 보상 요소로 선정하였으며, 구체적인 보상함수는 Table 5와 같다.

보상함수 #1은 목표 위치 추종에 관한 보상으로, 이전 시점과 현재 시점의 목표 위치까지의 거리 차로 정의된다. 에이전트가 이전 시점보다 목표 위치에 더 가까워지면 양의 보상을, 멀어지면 음의 보상을 부여한다. 이와 같이 거리의 절대값이 아니라 거리 변화량을 보상함수로 사용함으로써, 에이전트가 매 시점에서 목표 위치로 가까워지는 행동을 선택하도록 했다.

보상함수 #2는 급격한 가감속으로 인한 과도한 운동을 방지하기 위한 페널티 성격의 보상함수이다. 편대 전환 과정에서 에이전트가 급격한 가감속 운동을 수행할 경우, 에너지 소모량이 증가한다. 또한, 가감속에 따른 유체동역학적 특성이 변화하여, 편대 전환 시 에이전트 간 안전성 측면에서 불확실성을 높일 수 있다. 따라서, 속도 크기에 비례하는 페널티를 부여하여, 목표 위치 추종 과정에서 불필요하게 큰 속도 입력이 발생하지 않도록 제한한다. 여기서 0.01은 목표 위치 추종 보상에 비해 속도 페널티가 과도하게 지배적이지 않도록 설정한 가중치이다.

Table 6 Hyperparameter

Parameter	Value
Batch size	76,800
Buffer size	76,800
Beta	0.01
Epsilon	0.2
Gamma	0.99
Lamda	0.95
Time horizon	200
Number of hidden layer	1
Number of epoch	15
Hidden units	64
Learning rate	0.0005
Critic learning rate	0.0005
Number of mini-batch	1
Max step	10,000,000

보상함수 #3은 충돌 회피에 관한 보상으로, 리더와 에이전트 간 거리 또는 에이전트와 에이전트 간 거리가 각각 정의된 안전 반경보다 작아질 경우, 페널티를 부여하도록 설계하였다. 만약 에이전트 간 거리 또는 리더와의 거리가 안전 반경보다 작아지면 −50의 페널티를 부여하고, 그 외에는 0을 부여하였다. 페널티 부여를 통해 충돌 위험을 명확하게 구분, 리더 및 인접 에이전트와의 안전거리를 유지할 수 있도록 했다.

보상함수 #4는 최소 및 최대 운항 심도 이탈을 방지한다. 수중 운동체는 실제 운용 과정에서 허용할 수 있는 최소 운항 심도와 최대 운항 심도가 존재하며, 이 심도 범위 내에서 운항해야 한다. 본 연구에서는 최소 운항 심도는 20m, 최대 운항 심도는 300m로 가정하였다. 에이전트가 운용 심도 범위를 벗어나는 경우 페널티를 부여함으로써, 목표 위치로 이동하는 과정에서 운항 심도 내에서만 거동하도록 제약한다.

보상함수 #5는 편대 전환 성공 보상으로, 모든 에이전트가 각 목표 위치에 도착했을 때 양의 보상이 부여된다. 다만, 목표 위치 도착 여부를 판정하는데 허용 오차를 고려하는 것이 안정적인 정책 학습에 필요했고, 목표 위치로부터 가장 멀리 떨어진 에이전트의 오차가 1m 이내에 도달하면 1의 보상을 부여하였다. 이는 모든 에이전트가 목표 위치로부터 기준값 이하의 거리에 도달하였을 때, 성공한 것으로 간주한다. 따라서, 개별 에이전트의 단독적인 목표 도달이 아닌, 편대에 속한 모든 에이전트가 목표에 도달하는 것을 학습 목표로 두고 있음을 의미한다.

3.3 강화학습 모델링 타당성 평가

앞서 설계한 행동, 상태, 관측, 및 보상함수를 이용하여, MAPPO 알고리즘을 적용하기 위한 정책 학습 환경을 구성한 후, 평가하였다. 정책 학습에 사용된 MAPPO의 주요 하이퍼파라미터는 Table 6과 같다. 정책 평가는 총 12개의 편대 전환 시나리오에 대해 수행하였으며, 정책의 학습 안정성과 최종 성능을 확인하기 위해 누적 보상, 마지막 위치 오차, 성공률을 주요 지표로 사용하였다.

정책 학습 환경은 3.1절에서 정의한 편대 전환 환경을 병렬적으로 실행하는 구조로 구성하였다. 각 환경은 하나의 초기 편대 형상에서 목표 편대 형상으로 전환하는 독립적인 시나리오를 가리키며, 병렬 환경 실행을 통해 서로 다른 편대 전환으로부터 경험 표본을 동시에 수집할 수 있다. 정책 학습 환경에서 리더의 초기 위치는 (0, 0, 100)으로 설정하였고, 리더 속력은 1kts로 고정하였다. 또한 편대 전환은 리더가 종 방향으로 200m를 이동하는 동안 수행되도록 설정하였다. 따라서 본 연구의 정책 학습 환경은 리더 추종과 편대 전환을 동시에 수행해야 하는 협력 제어 문제를 반영한다.

각 에피소드는 편대 전환이 완료되거나, 최대 시간에 도달하거나, 충돌 또는 심도 위반과 같은 종료 조건이 발생할 때까지 진행되도록 구성하였다. 각 에피소드의 최대 길이는 1000 step으로 가정하였다. 편대 전환의 성공은 단순히 마지막 시점에 편대에 속한 에이전트와 리더의 거리 오차의 합계가 감소하는 것만으로 판정하지 않았다. 즉, 리더가 전환 거리(여기서, 200m)를 이동한 이후 모든 에이전트의 최종 위치가 각자 할당된 목표 위치로부터 성공 기준값인 1m 이내에 수렴했는지에 따라 판정하였다. 즉, 성공은 리더의 목표 지점 도착과 전체 편대에 속한 에이전트의 편대 전환이 동시에 완료되어야 함을 의미한다. 이를 terminated 상태로 분류하며, 그 외에는 truncated 상태로 분류, 처리하였다. 이것은 시간 초과에 의한 종료와 실제 과업 실패를 구분함으로써, 학습 과정에서 value bootstrap의 왜곡을 줄이고 정책 및 가치 함수 학습의 안정성을 높이기 위한 것이다(Qin et al., 2021). 리더–에이전트 간 충돌과 에이전트–에이전트 간 충돌은 모두 실패한 것으로 간주하였다. 충돌이 발생하면 해당 에피소드는 즉시 종료하였다.

운항 심도를 이탈한 때도 에피소드 실패, 종료 조건으로 고려하였으나, 정책 학습 과정에서는 심도 위반이 발생하더라도 즉시 종료하지 않고 패널티만 부여한 채 에피소드를 계속 진행하도록 하였다. 반면 정책 평가 과정에서는 심도 위반 발생 시 해당 에피소드를 즉시 종료하도록 하였다. 이러한 구성은 학습 단계에서는 탐색을 과도하게 억제하지 않으면서도, 평가 단계에서는 엄격한 심도 제약을 적용하기 위한 것이다.

정책의 학습 단계와 평가 단계에는 서로 다른 성공 기준값을 적용하였다. 학습 단계에서는 성공 기준값을 30m로 가정하여, 비교적 완화된 조건에서 학습의 안정성을 확보하고자 했다. 반면, 학습 환경에서는 에이전트–에이전트 간 최소 안전거리 기준값을

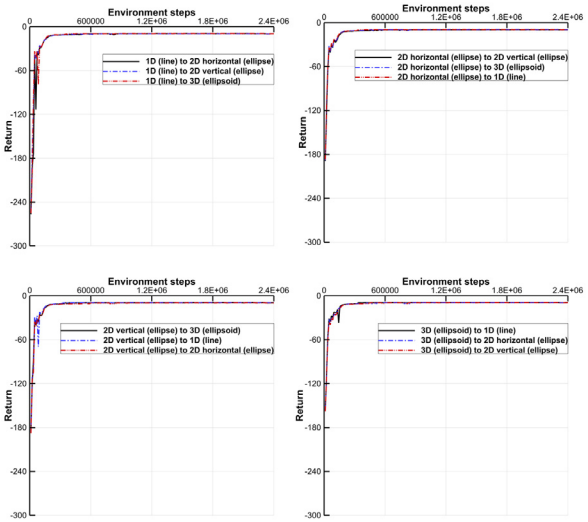

Fig. 6 Results of policy evaluation(return)

0.2m, 리더–에이전트 간 최소 안전거리 기준값을 0.5m로 각각 설정하였다. 여기서 리더–에이전트 간 최소 안전거리를 에이전트–에이전트 간 거리보다 크게 설정한 이유는, 유인 에이전트의 안전을 최우선으로 하기 위함이다. 정책의 평가 단계에서는 성공 기준값을 1.0m로 가정하였으며, 에이전트–에이전트 간 최소 안전거리 기준은 1.0m, 리더–에이전트 간 최소 안전거리 기준은 1.5m로 강화하였다. 따라서 평가 환경에서는 리더가 목표 위치에 도달하고, 동시에 모든 에이전트의 최종 위치 오차가 성공 기준값 이내에 수렴해야 편대 전환에 성공한 것으로 판정되며, 충돌에 대해서도 더 엄격한 기준이 적용된다. 따라서, 평가 단계에서는 실제 유·무인 복합 수중 운동체에 요구될 성능 조건을 고려한 정밀한 편대 전환 및 유지 성능을 만족하기 위함이다.

Fig. 6은 총 12개의 편대 전환에 대해 학습된 정책의 누적 보상 결과를 나타낸다. 모든 편대 전환 환경에서 누적 보상은 학습 초기에 크게 향상된 이후, 최종적으로 약 -9 내외로 수렴하는 경향을 보인다. 보상함수 #1은 목표 위치 추종에 따른 보상을 부여하고, 보상함수 #3, #4, #5는 각각 충돌, 심도 이탈, 편대 전환 성공 여부에 따라 부여된다. 보상함수 #2는 속도 크기에 비례하는 페널티로, 에이전트가 이동하는 시점마다 지속적으로 누적된다. 학습이 충분히 진행된 이후에는 충돌 페널티와 심도 이탈 페널티가 발생하지 않고, 목표 위치 추종 보상 또한 편대 전환 진행률이 높아질수록 점차 감소하므로, 누적 보상값은 보상함수 #2의 지배적인 영향을 받는다. 또한, 모든 편대 전환 환경에서 누적 보상이 유사한 값으로 수렴하는 것은 공평한 헝가리안 알고리즘 적용과 밀접한 관련이 있다. 각각의 에이전트가 목표 위치까지의 이동하는 데 있어, 균등한 이동 거리가 요구된다. 따라서, 각각의 에이전트에게 요구되는 이동 속도 또한 유사한 수준을 유지한다. 그 결과 보상함수 #2의 누적값이 유사하게 계산되었음을 알 수 있다.

Fig. 7은 학습한 정책에 대하여 각 에이전트의 위치로부터 부여된 목표 위치까지의 거리 오차를 가리킨다. 모든 편대 전환 환

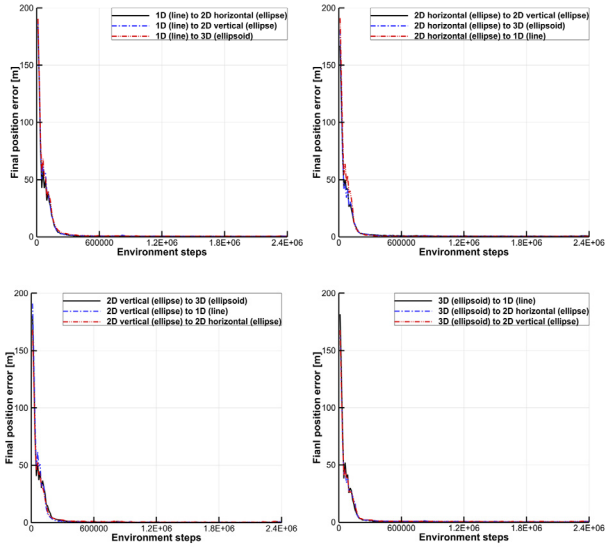

Fig. 7 Results of policy evaluation(final position error)

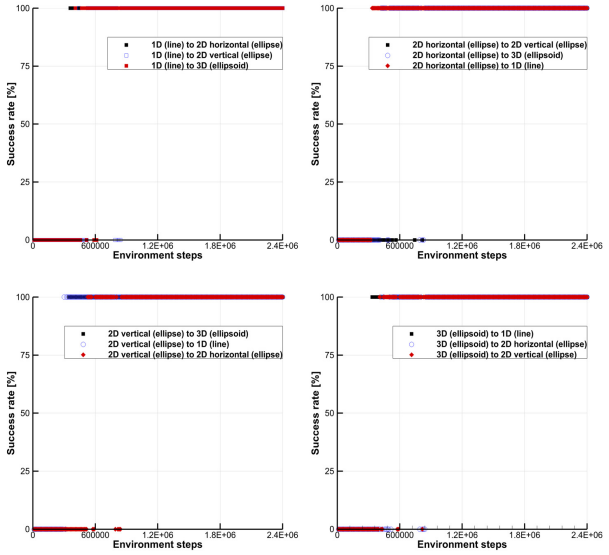

Fig. 8 Results of policy evaluation(success rate)

경에서 학습 초기에는 큰 오차를 보였으나, 학습이 진행됨에 따라 급격히 감소하여 최종적으로 0에 가까운 값으로 수렴하였다. 특히 초기 편대 형상별로 다른 3개 편대로 전환되는 경우의 마지막 위치 오차를 평균하여 비교하면 다음과 같다. 1차원 편대(직선)에서 다른 3개 편대로 전환될 때의 평균 거리 오차는 0.524m, 2차원 수평면 편대(타원형)에서 다른 3개 편대로 전환될 때의 평균 거리 오차는 0.580m, 2차원 수직면 편대(타원형)에서 다른 3개 편대로 전환될 때의 평균 거리 오차는 0.581m, 3차원 편대(타원체)에서 다른 3개 편대로 전환될 때의 평균 거리 오차는 0.520m로 나타났다. 모든 환경에서 최종 시점의 평균 오차가 약 0.5m 내외임을 통해 모두 편대 전환에 성공했음을 알 수 있다.

Fig. 8은 학습된 정책의 성공률을 평가한 결과를 나타낸다. 학습 초기에는 대부분의 환경에서 성공률이 0%에서 시작하지만, 학습이 진행됨에 따라 최종적으로 100%에 수렴하는 것을 확인할 수 있다. 학습된 정책이 완화된 학습 조건에만 적응한 것이 아니라 더욱 엄격한 평가 조건에서도 안정적으로 편대 전환을 성공적으로 수행하고 있음을 가리킨다.

앞서 기술한 12개의 편대 전환에 대해 Fig. 6, 7, 8을 종합적으로 검토한 결과, 누적 보상, 최종 시점의 위치 오차, 성공률의 변화 양상은 유사한 경향성을 보여준다. 즉, 제안된 강화학습 모델이 학습 횟수가 증가함에 따라 모든 값이 안정적으로 수렴하고, 성공적인 편대 전환이 이루어지고 있음을 가리키고 있는 것으로 판단된다.

4. 시뮬레이션 결과 및 분석

4.1 시뮬레이션 개요

제안된 MAPPO 기반 편대 전환 정책을 이용하여, 다양한 조건에서 유·무인 복합 수중 운동체의 편대 전환 시뮬레이션을 수행하였다. 학습된 정책이 정책 학습 환경과 동일한 조건에서 안정적으로 편대 전환을 수행할 수 있는지를 평가하는 동시에, 모든 편대 전환 시나리오에 대해 일관된 성능을 보이는지를 확인하였다.

시뮬레이션 조건은 정책 학습 환경과 같다. 즉, 하나의 편대가 총 12개의 에이전트로 구성되며, 2개의 층으로 구분된다. 내부 층에 4개, 외부 층에 8개의 에이전트가 배치된다. 리더의 초기 위치는 (0, 0, 100)으로 가정하였고, 편대 전환 거리는 200m로 정의하였다. 편대 전환 조건과 관계없이, 리더는 초기 위치로부터 종 방향으로 200m 이동하며, 각 에이전트는 목표 편대로 전환을 위하여 위치를 변경해야 한다. 시뮬레이션에서는 리더 속도 벡터를 (1, 0, 0)으로 설정하였다. 또한 전체 편대 전환 시나리오는 1차원(직선), 2차원 수평면(타원형), 2차원 수직면(타원형), 3차원(타원체) 편대가 차례대로 전환되도록 설정하였다. 전체 시나리오는 200m 길이의 개별 편대 전환 구간이 연속적으로 이어진 형태로 구성된다.

4.2 편대 전환 시뮬레이션 결과 및 분석

대표 시나리오로는 3차원 편대(타원체)에서 1차원 편대(직선)로 전환하는 경우를 선택하였다. 이 조건에서는 에이전트 간 상대 위치 변화가 가장 큰 사례이므로, 학습된 정책의 편대 전환 성능을 대표적으로 확인하기에 적절하다고 판단했다.

Fig. 9는 대표 시나리오에서 12개 에이전트의 이동 궤적을 나타낸다. 리더는 종 방향으로 전진하여 최종적으로 목표 지점인 (200, 0, 100)에 도달했다. 2개 층에 분포하는 모든 에이전트는 리더를 기준으로 목표 위치를 향해 이동하고 있음을 확인할 수 있다. 특히, 편대 전환 초기에는 모든 에이전트가 3차원 공간상에 분포하고 있었음을 확인할 수 있고, 편대 전환 후에는 횡 방향

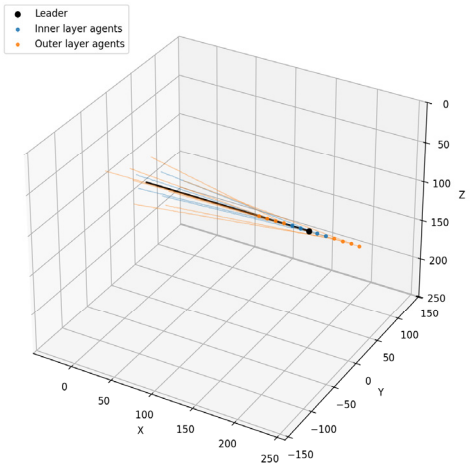

Fig. 9 Trajectory of formation transformation from 3D (ellipsoid) to 1D(line)–representative scenario

Table 7 Positions and errors of agent when changing formation from 3D(ellipsoid) to 1D(line)–representative scenario

Index	Start position	Target position	Final position	Error [%]
0	(19.84, 0, 111.25)	(220.00, 0.00, 100.00)	(219.86, 0.22, 100.58)	0.09
1	(−21.42, 13.73, 103.75)	(180.00, 0.00, 100.00)	(180.19, 0.14, 100.18)	0.09
2	(2.54, −20.26, 96.25)	(190.00, 0.00, 100.00)	(190.15, −0.07, 99.99)	0.07
3	(12.07, 11.02 88.75)	(210.00, 0.00, 100.00)	(210.02, 0.35, 100.36)	0.01
4	(24.21, 0.00, 121.88)	(240.00, 0.00, 100.00)	(239.88, 0.07, 100.34)	0.07
5	(−28.78, 18.46, 115.63)	(150.00, 0.00, 100.00)	(150.04, 0.37, 100.41)	0.02
6	(4.05, −32.32, 109.38)	(230.00, 0.00, 100.00)	(230.33, −0.24, 100.51)	0.12
7	(30.18, 27.56, 103.13)	(250.00, 0.00, 100.00)	(249.88, 0.43, 100.25)	0.08
8	(−48.85, −6.05, 96.88)	(140.00, 0.00, 100.00)	(140.02, 0.11, 100.18)	0.01
9	(39.11, −17.41, 90.63)	(260.00, 0.00, 100.00)	(259.78, −0.22, 100.21)	0.10
10	(−10.13, 26.39, 84.38)	(170.00, 0.00, 100.00)	(170.05, 0.38, 100.11)	0.00
11	(−11.16, −15.04, 78.13)	(160.00, 0.00, 100.00)	(159.91, 0.15, 100.41)	0.01

및 수직 방향 위치가 1차원 직선 형태로 변화했음을 확인할 수 있다. 따라서, 목표했던 편대 전환이 이루어졌음을 보여준다.

Table 7은 대표 시나리오에서 각 에이전트의 초기 위치, 목표 위치, 최종 위치 및 목표 위치 대비 최종 위치 오차율을 각각 정리한 것이다. 여기서 최종 위치 오차율은 초기 위치와 목표 위치

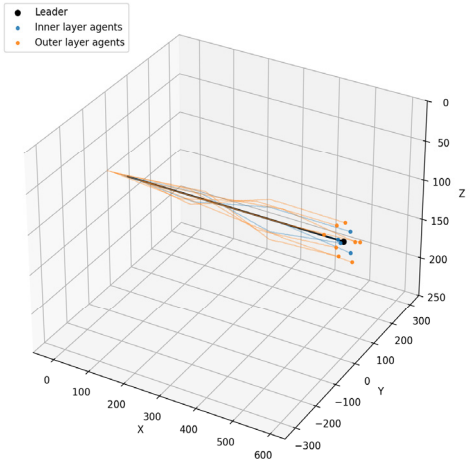

Fig. 10 Trajectory of full formation transformation

사이의 거리를 기준값으로 하고, 초기 위치와 최종 위치 사이의 거리와의 차이를 백분율로 환산하여 계산하였다. 그 결과, 전체 에이전트에 대한 평균 오차율은 0.06%, 최대 오차율은 0.12%로 나타났다. 또한 목표 위치와 최종 위치 사이의 거리 오차의 경우, 12개 에이전트의 평균 위치 오차는 0.12m, 최대 위치 오차는 0.27m로 나타났다. 모든 에이전트의 최종 위치 오차가 1m 이내로 본 연구에서 가정한 편대 전환 성공 조건을 달성했음을 확인할 수 있다. 모든 에이전트 최종 위치를 보면, 모든 에이전트가 1차원 편대(직선) 형상을 유지하고 있음을 확인할 수 있다. 기준 시나리오 결과는 학습된 MAPPO 기반 정책이 학습 환경과 동일한 조건에서 매우 안정적으로 동작함을 보여준다.

Fig. 10은 전체 편대 전환 시나리오에 대한 시뮬레이션 결과를 나타낸다. Fig. 10에서 확인할 수 있듯이, 리더를 중심으로 내부 층 및 외부 층 에이전트가 각 전환 구간에서 목표 편대 형상에 맞추어 재배치되며, 전체 전환 과정에서도 편대 구조를 유지하면서 안정적으로 이동하는 경향을 보였다. 이는 제안된 MAPPO 기반 편대 전환 정책이 단일 편대 전환뿐 아니라, 연속적으로 이루어지는 복합 편대 전환 상황에서도 적용할 수 있음을 보여준다. 특히 각 전환 구간이 독립적인 형상 변화임에도 불구하고, 전체 시나리오에서 에이전트의 궤적이 큰 불안정성 없이 연속적으로 연결된다는 점에서, 학습된 정책이 연속적인 편대 재구성 문제에 대해서도 일정 수준 이상의 강건성을 보유하고 있음을 확인할 수 있다.

5. 결 론

본 연구에서는 다중 에이전트 강화학습을 적용하여, 유·무인 복합 수중 운동체의 편대 전환 및 유지에 관한 프레임워크를 개발하고, 적용 가능성을 검토하고자 했다. 본 연구를 통해 다음과 같이 두 가지 결론을 얻을 수 있었다.

첫째, 유·무인 복합 수중 운동체의 편대 전환 문제를 모델링하기 위하여, 점질량으로 가정된 12개의 수중 운동체를 2개 층으로 나누어 배치하되, 기하학적으로 정의된 1차원, 2차원 수평면, 수

직면, 3차원 편대로 배치하는 프레임워크를 제안하였다. 통상적인 수중 운동체의 임무 수행에 필요한 유체동역학적 거동을 고려하여 선택한 것이며, 전략 자산의 가치에 따라 계층적 구조를 제안한 것이다.

둘째, 12개의 에이전트를 대상으로 다중 에이전트 강화학습을 적용하기 위하여, 행동, 관측, 상태 및 보상함수를 정식화하였다. 또한, 설계된 변수 및 함수를 적용하여, 편대 전환 정책이 효과적으로 학습될 수 있음을 확인하였다. 추가로, 학습 단계에서 완화된 성공 기준과 제약 조건을 적용함으로써 초기 탐색의 안정성을 확보했으며, 평가 단계에서는 엄격한 성공 기준과 제약 조건을 적용함으로써 높은 수준의 편대 전환 성능을 확보하고자 했다. 결과적으로, 공평한 헝가리안 알고리즘에 따라 12개 에이전트의 이동 거리와 관련된 누적 보상이 안정적으로 수렴했으며, 최종 위치에서의 오차는 1m 이내임을 확인할 수 있었다. 편대 전환 성공률 또한 100%에 수렴하는 것으로 나타났다.

본 연구는 유·무인 복합 수중 운동체 편대 전환을 위하여 다중 에이전트 강화학습을 적용하는 프레임워크 초기 개발에 관한 연구로서, 향후 수중 운동체의 6 자유도 유체동역학 수학모형을 고려한 유·무인 복합체계 운용 개념 연구의 기초 연구로 활용될 수 있을 것으로 기대된다. 다만, 본 연구는 유·무인 복합 수중 운동체 편대 전환을 위한 다중 에이전트 강화학습 기반 초기 프레임워크 개발에 목적을 둔 연구로서, 실제 수중 운동체의 동역학 특성을 모두 반영하지 못했다는 한계가 있다. 또한, 수중 통신의 제약, 해저 지형에 따른 장애 요소, 각 에이전트 측위 정보의 오차 등을 고려하지 못했다. 이외에도, 보상함수의 추가적인 고려, 에이전트 숫자, 편대 구성 등에 따른 일반화를 위한 후속 연구가 필요함을 인지하였다. 본 연구의 후속 연구로써 수중 운동체 6 자유도 운동 방정식을 적용한 강화학습 모델의 심화, 발전시킬 예정이며, 단계적 연구 계획 수립을 통해 언급된 한계점을 수정, 보완할 예정이다.

후 기

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2024-00395678, 우수신진연구). 또한, 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (RS-2025-02263945, 산업혁신인재성장지원사업).

Nomenclature

Abbreviations	Full meaning
CTDE	Centralized Training and Decentralized Execution
MAPPO	Multi-Agent Proximal Policy Optimization
MUM-T	Manned-Unmanned Teaming
NATO	North Atlantic Treaty Organization
PPO	Proximal Policy Optimization
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
USV	Unmanned Surface Vehicle

Symbols	Full meaning
$\hat{A}_t$	Estimated advantage used to update actor policy
B	Mini-batch sampled from rollout buffer
D	Rollout buffer storing collected trajectories
$\hat{E}_t$	Empirical expectation
$L^{CLIP}(\theta)$	Clipped surrogate objective for policy update
$L^{ENTROPY}(\theta)$	Entropy for encouraging exploration
$L^{VALUE}(\phi)$	Value loss function for critic update
$L(\theta,\phi)$	Total MAPPO objective function
N	Number of agents
R_{AA}	Safety radius between agents
R_F	Formation transition progress ratio
R_{LA}	Safety radius between leader and agent
$\hat{R}_t$	Cummulated rewards computed from the collected trajectory
V_i	Velocity of agent #i
V_i^j	Relative velocity of neighboring agent j with respect to agent #i
$V_\phi(s_t)$	State value estimated by critic at time step t
a	Action
c_1	Coefficient for value loss
c_2	Coefficient for entropy
d_i^{Goal}	Distance from agent #i to goal position
d_i^L	Distance form agent #i to leader
d_i^j	Distance from agent #i to neighboring agent #j
$e_{x,y}$	Distance between the initial position and the final position
i	agent index
k	Policy update iteration index
o	Observation
p_L	Position of leader
p_{goal}	Position of goal position
p_i	Position of agent
p_j	Final position of agent
p_i^j	Relative position of neighboring agent #j with respect to agent #i
r	Reward
$r_t(\theta)$	Probability ratio between current policy and old policy at time step t
s	State
z	Depth
$z_{\max}$	Maximum allowable operating depth

z_{min}	Minimum allowable operating depth
ϵ	Clipping parameter
θ	Decentralized actor parameters
θ_{old}	Old actor parameters used for PPO update
π	Policy
ϕ	Centralized critic parameters
$\varnothing$	Empty set
$\sim$	Sampled from
$\subset$	Subset of

References

Cao, F., Xu, H., Ru, J., Li, Z., Zhang, H., and Liu, H., 2025. Collision Avoidance of Multi-UUV Systems Based on Deep Reinforcement Learning in Complex Marine Environments. *Journal of Marine Science and Engineering*, 13(9).

Gorji, S., and Granmo, O., 2023. Off-policy and on-policy reinforcement learning with the Tsetlin machine. *Applied Intelligence*, 53, pp.8596-8613.

Hou, X., Guo, Z., Wang, X., Qian, T., Zhang, J., Qi, S., and Xiao, J., 2022. Parallel learner: A practical deep reinforcement learning framework for multi-scenario games. *Knowledge-Based Systems,* 25.

Islam K., Ahmad, I., Habibi, D., and Waqar, A., 2022. A Survey on energy efficiency in underwater wireless communications. *Journal of Network and Computer Applications*, 198.

Lee, J., 2022. An Analysis of the Evolution of War and the Changes of War Patterns. *Korea Maritime Security Review*, 5(2), pp.105-131.

Liu, Z., Cai, W., and Zhang, M., 2023. Reinforcement Learning-based path tracking for underactuated UUV under intermittent communication. *Ocean Engineering*, 288(1).

Mead, T., Wang, Z., Foo, E., Dong, J., Dong, N., Ko, R., Koay, A., Nguyen, K., Xu, Y., Kim, J., Bornstein, S., and Hou, Z., 2026. Multi-agent reinforcement curriculum learning for real unmanned ground vehicles. *Engineering Applications of Artificial Intelligence*, 167.

Moon, s., 2022. Fair Hungarian Algorithm for Swarming Drone Flight Formation Transformation. *Journal of Korean Institute of Information Scientists and Engineers*, 49(6), pp.459-465.

Ministry of National Defense, 2023. *Defense Innovation 4.0*.

Nam, S., Cho, Y., and Woo, J., 2023. Reinforcement Learning for Minimizing Tradiness and Set-Up Change in Parallel Machine Scheduling Problems for Profile Shops in Shipyard. *Journal of the Society of Naval Architects of Korea*, 60(3), pp.202-211.

NATO, 2021. *Artificial Intelligence and Autonomy in the Military: An Overview of NATO Member States' Strategies and Deployment*.

NATO, 2025. *MASTERING THE FUTURE OF UNCREWED WARFARE*. NATO Report No.023 STCTTS 25 E.

Ning, Z., and Xie, L., 2024. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3, pp.73-91.

Qin, G., Luo, Q., Yin, Y., Sun, J., and Ye, J., 2021. Optimizing matching time intervals for ride-hailing services using reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 129.

Shin, D., Park, B., Lim, C., Oh, S., Kim, G., and Shin, S., 2020. Pipe Routing using Reinforcement Learning on Initial Design Stage. *Journal of the Society of Naval Architects of Korea*, 57(4), pp.191-197.

Shixin, Z., Feng, P., Anni, J., Hao, Z., and Qiuqi, G., 2024. The unmanned vehicle on-ramp merging model based on AM-MAPPO algorithm. *scientific reports*, 19416(14).

Sutton, R., and Barto, A., 2018. *Reinforcement Learning: An Introduction*. The MIT Press.

Tan, M., Sun, H., Zhou, H., Leng, Z., and Ding, D., 2026. Biased random masked attention MAPPO algorithm for zero-shot scale generalization of multi-UAV air combat. *Journal of Computational Design and Engineering*, 13, pp.46-68.

Tian, S., Yang, M., Xiong, R., He, X., and Rajasegarar, S., 2024. A sequential multi-agent reinforcement learning framework for different action spaces. *Expert Systems with Applications*, 258.

U.S. ARMY, 2010. *UNMANNED AIRFRAFT SYSTEMS ROADMAP 2010-2035.*

U.S. DEPARTMENT OF DEFENSE, 2011. *UNMANNED SYSTEMS INTEGRATED ROADMAP*.

Xing, X., Zhou, Z., Li, Y., Xiao, B., and Xun, Y., 2024. Multi-UAV Adaptive Cooperative Formation Trajectory Planning Based on an Improved MATD3 Algorithm of Deep Reinforcement Learning. *IEEE TRANSATIONS ON VEHICULAR TECHNOLOGY*, 73(9), pp.12484-12499.

Zhang, D., Yuan, Q., Meng, L., Xia, R., Liu, W., and Qin, C., 2026. Reinforcement learning for single-agent to multi-agent systems: from basic theory to industrial application progress, a survey. *Artificial Intelligence Review*, 59(46).

Zhang, H., and Yu, T., 2020. *Deep Reinforcement Learning*. Springer.

Zhao, Y., Ma, Y., and Hu, S., 2021. USV Formation and Path-Following Control via Deep Reinforcement Learning With Random Braking. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 32(12), pp.5468-5478.

Authorship Contribution Statement

Seungtae Kim: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draftungjun You; **Youngjun You:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing - review & editing

김 승 태

유 영 준