JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 11, pp. 79-88, November 2024
https://doi.org/10.9708/jksci.2024.29.11.079

Multi-Agent Reinforcement Learning based Swarm Drone using QPLEX and PER

Jin-Ho Ahn*, Byung-In Choi*, Tae-Young Lee*, Hae-Moon Kim*, Hyun-Hak Kim*

*Engineer, Intelligence Software Team, Hanwha Systems Co., Ltd., Korea

[Abstract]

With the advancement of unmanned aerial vehicle technology, swarm drones are increasingly being deployed across various domains, including disaster response and military operations, where their effectiveness is particularly pronounced. These swarm drones leverage real-time data sharing and decision-making capabilities to execute tactical missions, making collaborative behavior essential in complex battlefield environments. However, traditional rule-based behavior mechanisms face limitations as environmental complexity escalates. This paper explores the potential of applying multi-agent reinforcement learning (MARL) to swarm drone models and proposes strategies to enhance their mission success rates. By utilizing QPLEX and Prioritized Experience Replay (PER), we present methods aimed at improving learning efficiency. Validation through the SMACv2 simulator reveals that the proposed approach achieves faster learning convergence and higher mission success rates compared to existing MARL algorithms.

▸ **Key words**: Multi-agent reinforcemet learning, swarm drones, QPLEX, PER, SMAC

[요 약]

무인 항공기 기술의 발전으로 군집 드론은 재난 구조 및 군사 작전 등 다양한 분야에서 활용되고 있으며, 특히 군사 작전에서 그 효과가 두드러진다. 군집 드론은 실시간 데이터 공유와 의사결정 능력을 통해 효과적인 전술 작전을 수행할 수 있어, 복잡한 전장 환경에서의 협업 능력이 필수적이다. 그러나 기존의 규칙 기반 행동 매커니즘은 환경의 복잡성이 증가함에 따라 한계를 보인다. 이에 본 논문에서는 군집 드론 모델에 다중 에이전트 강화학습 적용 가능성을 검토하고 군집 드론의 임무 성공률을 향상시키기 위한 방안을 제안한다. QPLEX와 PER를 활용하여 학습 효율을 높이는 방법을 제안하였으며, SMACv2 시뮬레이터를 통해 검증한 결과, 제안된 방법은 기존 MARL 알고리즘보다 빠른 학습 수렴 속도와 높은 임무 성공률을 기록하였다.

▸ **주제어**: 멀티 에이전트 강화학습, 군집 드론, QPLEX, PER, SMAC

• First Author: Jin-Ho Ahn, Corresponding Author: Byung-In Choi
*Jin-Ho Ahn (babupeople@hanwha.com), Intelligence Software Team, Hanwha Systems Co., Ltd.
*Byung-In Choi (byungin.choi@hanwha.com), Intelligence Software Team, Hanwha Systems Co., Ltd.
*Tae-Young Lee (ty.lee@hanwha.com), Intelligence Software Team, Hanwha Systems Co., Ltd.
*Hae-Moon Kim (haemoon1205@hanwha.com), Intelligence Software Team, Hanwha Systems Co., Ltd.
*Hyun-Hak Kim (kim.hyun.hak95@hanwha.com), Intelligence Software Team, Hanwha Systems Co., Ltd.
• Received: 2024. 10. 11, Revised: 2024. 10. 25, Accepted: 2024. 11. 01.

Copyright © 2024 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

드론 활용이 재난 구조, 환경 모니터링 등의 다양한 분야에서 활용성이 증대되고 있다[1]. 최근에는 그 활용성이 군사 작전 수행에서 두드러진다. 특히 러시아-우크라이나 전쟁 간 저가의 드론 활용을 통해 주요 전략 자산을 파괴하거나, 레이더 감시를 피해 적 상황을 정찰하는 등의 공격 및 정찰의 목적으로 드론을 활용하므로 국지전 승리를 끌어내는 장면들을 여러 매체를 통해 확인되고 있다. 이렇듯 드론은 군사 분야에서 전쟁 승패를 결정할 만큼 중요도가 높아졌다. 또한, 드론 기반 임무 성공률 향상을 위해 단일 드론보단 군집 드론 운용을 통한 작전 수행이 필요하다[2].

군집 드론 운용 복잡도는 개체 수의 증가와 협업 임무 여부 등 다양한 요인에 의해 좌우된다. 즉 군집 드론의 활용은 단일 드론 간 협업을 통해 공통 목적을 달성해야 하므로 불리한 전장 상황에서도 개별 환경의 인식 및 공유가 필요하며, 재밍 같은 예측 불가능한 변수에 대응할 수 있어야 한다. 예를 들어, 다수의 적 개체를 유인해 섬멸하거나 전술적으로 불리한 위치를 인식하고 회피하는 능력은 군집 드론이 성공적으로 임무를 수행하는 데 필수적이다. 이러한 요구를 충족시키기 위해서는 드론 간의 임무 할당을 통한 역할 분배와 변칙적인 상황에 대처 가능한 군집 전술 알고리즘이 필요하다.

기존 군집 드론은 현재 상태마다 행동을 정의한 규칙 기반 행동 메커니즘을 사용하였다[3]. 하지만 동일 임무에 대한 협업, 다중 임무 수행을 위한 개별 목표 수립 등을 고려했을 경우 기존 규칙 기반 제어 방식은 환경의 복잡도가 높아질수록 규칙 부여의 복잡성 및 주변 환경 변화 대처의 한계점을 보였다. 이런 문제점을 해결하고자 강화학습(Reinforcement Learning)을 적용하여 복잡한 상황에 대한 동적 대처 능력을 향상하고자 하였으나 드론 간 통신 연결, 센서 제어 등의 문제로 개별 상황인식 결과의 공유가 어려워 군집 임무 수행에 실패했다. 또한, 강화학습 적용은 단일 에이전트 강화학습으로 드론 자세 제어, 경로 계획[4] 등에 치중되어 있어 다수 개체 제어 알고리즘으로 적합하지 않았다. 하지만 반도체 발달과 함께 센서 데이터 획득 안정성, 임베디드 보드 처리 능력이 향상되었고 5G, 6G 통신을 통해 개체 간 안정적인 통신이 가능해졌다. 이러한 하드웨어의 발전으로 인해 전술 모델 학습으로 MARL(Multi Agent Reinforcement Learning)을 적용하는 연구가 진행되고 있다[5].

MARL은 개별 에이전트 간 협력을 통해 단일 또는 다중 임무를 수행할 수 있는 모델 학습 방법으로 드론 군집 전술 모델에 효과적으로 적용될 수 있다. 전장 상황에서 각 드론은 독립적으로 상황을 판단하며 다른 드론과의 상호작용을 통해 최적의 경로 수립을 통해, 적의 병력 회피 후 임무 수행을 하는 복잡한 작업을 수행할 수 있다. MARL 알고리즘의 활용은 드론 간의 협업 임무를 계획하고 실시간으로 변화하는 전장 상황에 대해 유연하게 대응할 수 있게 한다. 이런 장점에도 군집 드론 제어를 위해 MARL 적용 시, 반드시 고려해야 할 사항은 학습 효율이다. 군집 드론의 임무는 복잡하고 동적인 환경 속에서 수행되기 때문에, 학습 알고리즘의 선택이 임무 성공률에 직접적인 영향을 미친다. 특히, 다양한 전술적 상황의 가정에 따라, 학습을 통한 제어 알고리즘 검증을 위해선 신속한 학습과 함께 높은 성능을 유지해야 한다. 따라서, 군집 제어 성능의 안정성을 고려한 알고리즘의 선정 및 학습 속도 개선에 대한 연구가 필요하다.

본 논문에서는 MARL 알고리즘 중 빠른 학습 수렴 속도를 보여주는 QPLEX(Duplex dueling multi-agent Q-Learning)[6]와 학습 데이터 가중치 부를 통해 학습 효율성을 향상한 PER (Prioritized Experience Replay) [7]를 제안한다. 제안된 방법으로 학습된 모델은 교전 시뮬레이터인 SMAC(Starcraft Multi Agent Challenge)[8]을 통해 가능성을 검증하고 기존 MARL 알고리즘과 비교 분석을 통해 제안된 방법의 성능 평가를 진행한다.

본 논문의 구성은 다음과 같다.

II장에서 학습 시뮬레이터인 SMAC을 소개하고 학습에 사용될 MARL과 PER를 설명한다. III장에서는 군집 드론 임무에 대해 정의하고 학습 네트워크 구성과 학습 파라미터를 설명한다. IV장에서는 제안한 알고리즘과 기존 알고리즘별 학습 결과를 누적 보상과 임무 성공률을 분석하여 군집 드론의 적합한 모델에 대한 평가를 진행한다. V장에서는 제안된 학습 모델 활용과 향후 추가적인 연구 방향에 대해 정리한다.

II. Preliminaries

MARL은 복잡한 환경에서 효과적인 의사결정을 지원하기 위해 Dec-POMDP(Decentralized partially observe Markov decision process)모델을 사용한다. 싱글 에이전트 강화학습에서 사용하는 MDP와 추가적으로 각 에이전트가 부분적으로 관찰 가능한 POMDP를 일반화한 것으로 $G = (s, \{a_i\}, \mathrm{T}, \mathrm{R}, \{o_i\}, \mathrm{O}, \gamma)$ 정의된다. s는 모든 에이전트들의 상태 집합으로 시나리오가 진행됨에 따라 변화

한다. 에이전트 i 는 조건부 관찰 확률 $O = P(o|s,a)$에 따라 부분 관찰값인 o_i를 획득할 수 있으며 o_i기반으로 에이전트는 행동을 결정한다. 이때, 에이전트는 정의된 행동 공간 a_i에서 행동 선택이 가능하다. 다음 스텝의 상태 s'는 모든 에이전트의 행동 $\boldsymbol{a}$, 상태값 s 기반 상태 전이 확률 모델 $T = P(s'|s,\boldsymbol{a})$에 의해 계산된다. R은 에이전트 행동에 대한 보상 함수이며 γ는 감쇠율이다.

멀티 에이전트 모델의 각 에이전트는 부분 관측값을 타 에이전트에게 공유하고 상호작용을 통해 문제를 해결해야 한다. 따라서 전체 상태를 보고 판단할 수 있는 학습 네트워크 구조가 필수적이다. CTDE (Centralized Training with Decentralized execution)는 해당 사항을 만족할 수 있는 개념으로 학습 시 에이전트 전체 상태를 활용하여 네트워크를 업데이트하고 추론 시 각 에이전트의 부분 관측값 기반으로 행동을 도출한다. CTDE를 적용한 대표적인 알고리즘으로 다음과 같다.

MAPPO[9]는 싱글 에이전트 강화학습 알고리즘 중 PPO를 MARL에 적용한 것으로 actor-critic 네트워크 구조를 가지며, 과거 학습 데이터를 중요도 샘플링을 통해 현재 모델에 적합하게 재가공해 모델을 학습하여 학습 샘플 효율성이 높다.

VDN[10]은 결합 행동-가치함수 Q_{tot}를 각 에이전트의 행동 가치함수 Q_i로 분해가 가능하다는 가정을 한다. 에이전트별 행동에 따른 행동 가치함수를 분해함으로써 각 에이전트가 독립적으로 학습하는 것보다 빠르게 수렴하는 것을 입증했다.

$$Q_{tot}(s,a) =: \sum_{i=1}^{n} Q_i(s_i,a_i) \tag{1}$$

QMIX[11]의 핵심은 결합 행동-가치함수 Q_{tot}을 개별 에이전트의 행동 가치함수 Q_i의 조합으로 표현하는데 있다. 식(2)는 학습 네트워크의 단조성 증가 조건으로 해당 조건을 만족하면 더 복잡한 관계를 모델링 할 수 있다는 것을 전제로 한다. QMIX는 mixing network를 통해 분해된 에이전트의 행동 가치함수와 상태 정보기반으로 에이전트 간의 상호작용을 할 수 있게 유도함으로써 에이전트 간 협업 능력을 향상했다.

$$\frac{\partial Q_{tot}}{\partial Q_i} \geq 0,\ \forall i \in N \tag{2}$$

1. QPLEX

QQPLEX는 에이전트 행동 가치함수를 분해한 것과 mixing network를 이용한 부분에서 QMIX 구조와 동일하다. 하지만 QMIX에서 사용하는 행동 가치함수 Q를 이점함수 A와 상태 가치함수 V로 분해하여 mixing network를 학습한다는 점에서 차이가 있다.

$$\begin{aligned} Q_i(\tau_i,a_i) &= V_i(\tau_i) + A_i(\tau_i,a_i) \\ V_i(\tau_i) &= \max_{a_i'} Q_i(\tau_i,a_i') \end{aligned} \tag{3}$$

행동-관측 집합 τ_i와 다음 스탭 행동 가치 함수값을 최대로 하는 행동 a_i'를 통해 상태 가치함수를 구할 수 있으며 이점 함수 A_i는 Q_i, V_i의 차를 통해 계산된다. A_i는 에이전트가 선택한 행동이 현재 상태 기준으로 얼마나 좋은 선택이였는지 수치상으로 나타낼 수 있으며 QPLEX mixing network의 입력으로 사용된다. 에이전트별 A_i, V_i는 mixing network에서 합 연산을 통해 전체 에이전트의 행동 가치함수 Q_{tot}을 계산하고 이를 통해 손실 값을 구하는 데 사용된다.

$$\begin{aligned} Q_{tot}(\tau,a) &= V_{tot}(\tau) + A_{tot}(\tau,a) \\ V_{tot} &= \sum_{i=1}^{n} V_i(\tau) \\ A_{tot}(\tau,a) &= \sum_{i=1}^{n} \lambda_i(\tau,a) A_i(\tau,a_i) \end{aligned} \tag{4}$$

$\lambda_i(\tau,a)$는 MHA(multi-head attention) 모듈로 시그모이드함수를 사용한 MLP를 개별적으로 사용하여 key, agent, action값을 예측한다. 해당 값은 A_i의 중요도에 대한 가중치를 계산하며 이를 이용해 에이전트 선택의 일관성을 유지하고 신용할당 문제를 만족시켜 효율적인 학습을 가능하게 한다[12].

2. PER

강화학습에서 replay buffer는 샘플 효율성을 높이고, 연속적인 경험의 상관관계를 줄여 학습 안정성을 향상시키는 중요한 역할을 한다[14]. 에이전트가 환경과 상호작용하며 수집한 학습 데이터를 저장하고 무작위로 샘플을 선택하여 학습함으로써, 다양한 상태와 행동의 조합을 활용할 수 있다. PER는 기존 replay buffer에 저장된 학습 데이터에 우선순위를 추가로 부여한다. 우선순위는 해당 데이터의 TD loss와 사용 빈도수로 계산되며 각 데이터는 샘플링될 확률 p_i를 가지고 있다. 식(5)에 따라 우선순위 $P(i)$가 결정되며 p_i는 사용되는 시점에 갱신된다. 이때,

$$P(i) = \frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}} \tag{5}$$

α는 우선순위 지정의 양을 결정하는 하이퍼 파라미터로 0으로 수렴할수록 데이터가 균일하게 샘플링된다. 또, PER는 TD loss에 각 데이터의 가중치를 사용하는 중요도 샘플링을 적용함으로써 특정 데이터로만 학습했을 때 생길 수 있는 편향을 보정한다. 가중치 ω_i는 식(6)과 같이 학습 데이터 총수 N과 샘플링될 확률을 우선순위 $P(i)$를 통해 계산되며 β는 가중치 적용 정도에 대한 파라미터이다.

$$\omega_i = (\frac{1}{N} \cdot \frac{1}{P(i)})^{\beta} \quad (6)$$

PER는 데이터 우선순위와 중요도 샘플링을 통해 보다 가치 있는 학습 데이터를 먼저 학습하게 하며 해당 데이터 가치를 고려해 모델에 적용함으로써 학습 안정도와 샘플 효율성을 높인다.

3. SMACv2

SMAC은 MARL 연구를 위한 벤치마크 환경으로 블리자드사의 스타크래프트2 게임을 기반으로 개발되었다. 다양한 협력적 및 경쟁적 상황에서 에이전트가 상호작용을 하는 복잡한 시나리오를 제공하여 에이전트의 학습 및 전략적 의사결정 능력을 평가하는데 유용한 시뮬레이터이다. 최근 많은 연구에서 벤치마크로 사용되었지만, MARL 알고리즘의 지속적인 개선으로 인해 대부분의 시나리오에서 완벽한 승률을 기록하고 있어 더 이상 알고리즘 개선을 검증할 수 없다[13]. 이를 개선하기 위해 제안된 것이 SMACv2이다. 에피소드마다 팀 구성과 에이전트 시작 위치를 무작위로 생성하여, 에이전트가 고정된 시나리오를 반복하는 것이 아니라 다양한 시나리오 맞춰 학습할 수 있도록 환경을 제공한다. 기존 SMAC에서 제공된 시나리오는 유닛 수, 종류, 시작 위치가 고정되어 있어 사용자가 환경에 직접적으로 관여하지 못하는 형태였다. SMACv2의 경우 Fig. 1과 같이 시나리오 시작 시 유닛의 종류를 선택할 수 있고 유닛의 초기 위치를 변경하는 기능을 제공한다. 이에 따라 MARL 모델 학습 시 사용자가 원하는 환경을 조성할 수 있고 환경의 복잡도를 증가시켜 모델의 정확한 성능 검증을 할 수 있다.

III. The Proposed Scheme

1. Scenario description

본 논문에서는 SMACv2 시뮬레이터 엔진을 이용하여 군집 드론 모델을 학습한다. 에이전트에게 절대적으로 불리한 상황을 조성함으로써 학습된 모델의 정확한 성능 검증과 학습 알고리즘의 신뢰도를 높이는 데 초점을 맞춰 시나리오를 구성하였다. 시나리오의 조건은 다음과 같다.

Fig. 1. SMACv2 Scenario example

- ✓ 동일 무장/내구도, 행동 영역
- ✓ 아군 드론 수 < 적군 드론 수
- ✓ 포위된 환경

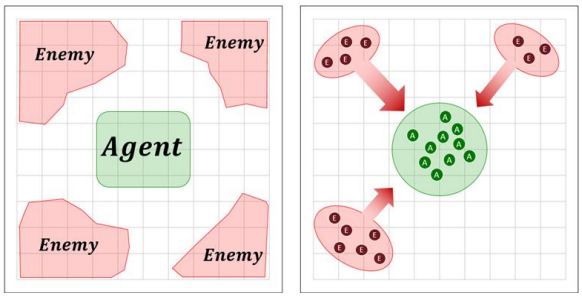

Fig. 2. Initial Positions of Entities in the Scenario

학습 시나리오 시작 시, 학습 대상인 아군 드론은 Fig.2(좌)의 빨간색 영역에서 밀집대형으로 생성되며 적군 드론은 파란색 영역 4곳 중 3곳을 랜덤하게 선택하고 해당 구역에 1기 이상의 적군 드론을 배치한다. Fig.2(우)는 적군 드론이 4기, 6기, 3기로 초기에 배치되어 아군 드론을 공격하는 시나리오 예시이다. 적군 드론은 아군 드론의 전멸을 목적으로 하며 아군 드론의 위치를 항상 제공받을 수 있다. 반면 아군 드론은 적군 드론의 위치를 제한된 범위에서 식별할 수 있다. 시나리오 목표는 적군 섬멸로 적군 드론을 모두 파괴해야 승리할 수 있다.

2. Learning parameter

2.1 observation & state

드론은 자신의 위치 기준 일정 거리 내에서만 아군 드론과 적군 드론 정보를 받을 수 있다. 이는 Dec-POMDP에

의해 각 에이전트는 부분 관측만 가능할 수 있는 가정을 지키기 위함으로 식별되지 않은 아군 드론은 협업할 수 없다. 또한 적군 드론을 발견하지 못하거나 공격 사거리 밖으로 식별된다면 공격이 불가하다. 아군 드론이 획득할 수 있는 관측값(observation)은 Table 1에서 확인할 수 있다.

상태값(state)은 CTDE개념을 적용한 MARL 알고리즘에서 필요한 입력 파라미터로 시나리오 내 모든 아군 드론의 관측값 집합을 의미한다. 상태값 수집 시 관측값 데이터 중 위치 데이터는 절대좌표계로 변환되는데 학습 시 정확한 에이전트 행동 평가를 하기 위함이다.

Table 1. observation

item	sub-item	type
own	movable direction (west, east, north, south)	bool
	health	float
	absolute position(x, y)	float
	weapon reload time	float
	unit type	one-hot
ally	identifiable	bool
	health	float
	relative distance	float
	relative position(x, y)	float
	unit type	one-hot
enemy	could attack	bool
	health	float
	relative distance	float
	relative position(x, y)	float
	unit type	one-hot

2.2 action

모델의 행동 공간(action space)은 discrete action 방식을 따르며 드론이 선택할 수 있는 행동(actioin)은 기동과 공격으로 분류할 수 있다. 기동은 [go west, go east, go north, go south]가 가능하며 시나리오 내 모든 드론은 피아 관계없이 서로 충돌할 수 없고 시나리오에서 제공하는 맵 밖으로는 나갈 수 없다. 아군 드론은 식별된 적군 드론이 공격 사거리 내 위치할 경우에 한정해서 공격 목표를 지정할 수 있다.

3. network architecture

본 논문에서는 모델의 임무 성공률과 데이터 학습 효율을 높이기 위해 MARL 알고리즘 QPLEX와 우선 순위기반 학습 데이터 가중치 적용기법인 PER를 제안한다. Fig.5는 제안된 알고리즘을 적용한 학습 구조를 도식화한 것으로 학습 데이터 생성, PER 기반 배치 데이터 샘플링, QPLEX 학습 구조를 이용한 TD loss 계산 과정을 나타낸다. 학습 네트워크는 크게 Duplex dueling 구조를 가지는 mixing network와 MLP, GRU로 구성된 agent network로 구분된다.

3.1 agent network

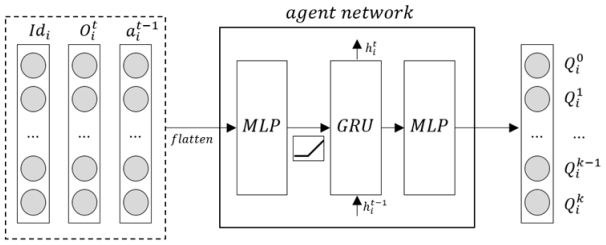

Fig. 3. agent network architecture

agent network는 에이전트의 행동 가치Q_i^k를 계산하며 $\epsilon-greedy$정책에 따라 행동 가치 Q가 가장 높은 행동을 식(7)에 따라 선택한다. 또한 각 에이전트는 모두 동일한 agent network를 사용하는 weight sharing 기법을 적용하여 학습 파라미터를 줄이고 과적합을 방지하였다. 이는 여러 에이전트 간의 행동이 일관되게 유지할 수 있게 도와주며 학습 속도, 파라미터 효율성을 높일 수 있다.

$$Q_i(\tau_i,a_i)=\max(Q_i^k(\tau_i,a_i)),\ \forall k\in a_i \tag{7}$$

3.2 mixing network

mixing network는 최종 결과값은 모든 에이전트의 행동 가치가 결합된 값으로 Transformation 모듈, Dueling mixing 모듈을 통해 순차적으로 계산되어 출력된다.

에이전트가 선택한 행동 Q_i는 V_i, A_i로 분해되어 mixing network의 Transformation 모듈에서 전체 상태값 s_t를 반영해 식(8)에 따라 계산된다.

$$\begin{aligned} Q_i(\boldsymbol{\tau},a_i)&=\omega_i(\boldsymbol{\tau})Q_i(\tau_i,a_i)+b_i(\boldsymbol{\tau})\\ V_i(\boldsymbol{\tau})&=\omega_i(\boldsymbol{\tau})V_i(\tau_i)+b_i(\boldsymbol{\tau})\\ A_i(\boldsymbol{\tau},a_i)&=Q_i(\boldsymbol{\tau},a_i)-V_i(\boldsymbol{\tau})=\omega_i(\boldsymbol{\tau})A_i(\tau_i,a_i) \end{aligned} \tag{8}$$

Transformation 모듈의 hyper network는 ω_i, b_i를 계산하는 네트워크로 구성되며 각각 2 layer MLP로 설계했다. 식(8)의 $\boldsymbol{\tau}$는 전체 상태값 s_t를 이용하며 hyper network의 활성화함수는 Relu를 사용하였다. 이러한 구성으로 $\omega_i(\boldsymbol{\tau})>0$ 조건을 만족시켜 에이전트의 탐욕적 행동 선택의 일관성을 유지하고 부분 관측 가능성을 완화하였다. Dueling Mixing 모듈은 key, agent, action으로 구성된 hyper network로 이루어져 있으며 에이전트별로 계산된 $V_i(\boldsymbol{\tau})$, $A_i(\boldsymbol{\tau},a_i)$, $\lambda_i(\boldsymbol{\tau},a)$기반으로 식(3)에 의해

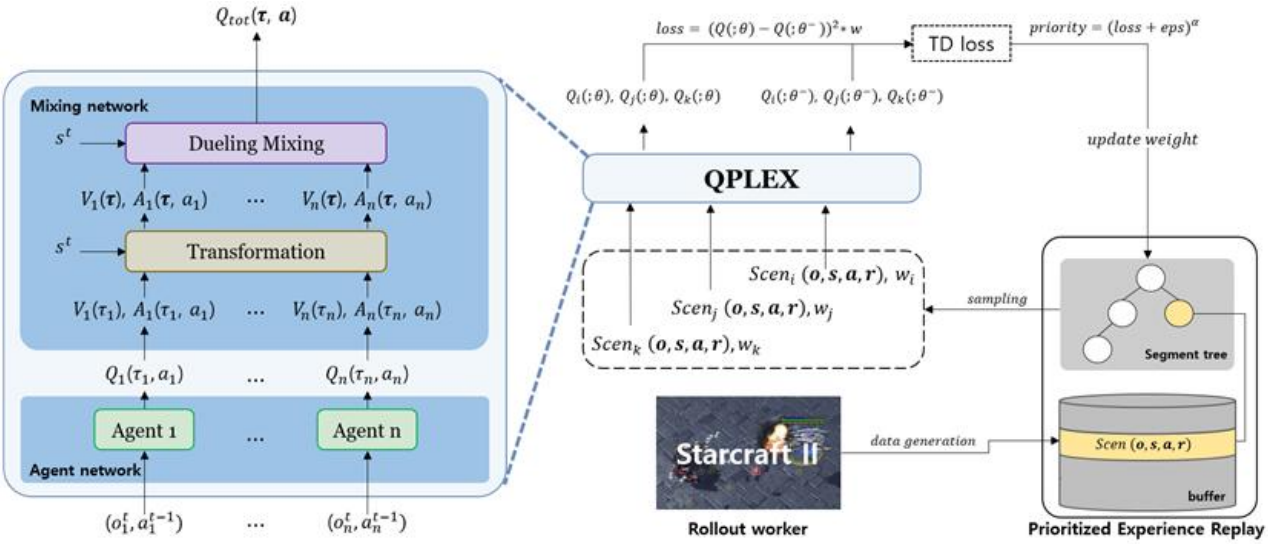

Fig. 4. proposed learning architecture

$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a})$를 계산한다. 이렇게 계산된 $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a})$은 모델 업데이트를 위한 TD loss를 구하는 데 사용된다.

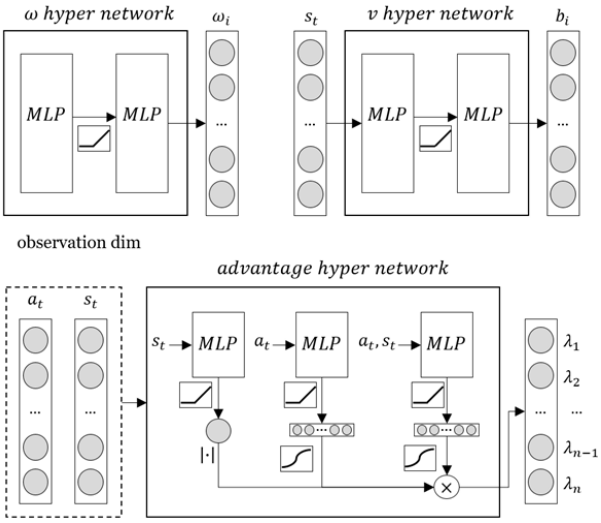

Fig. 5. mixing network architecture

4. Learning architecture

4.1 management of learning data

Rollout worker는 군집 드론 시나리오를 진행하면서 학습 데이터를 생성하는 역할을 하며 학습 데이터는 시나리오 시작부터 종료까지 모든 스텝을 하나의 배치 데이터로 저장한다. 배치 데이터는 PER 내부 버퍼에 저장되어 초기 가중치를 부여받고 Segment tree에 의해 관리된다. egment tree는 노드 구간 합을 이용해 우선순위를 효율적으로 계산할 수 있는 자료구조로 새로운 배치 데이터가 추가되거나 기존 배치 데이터의 우선순위가 변경되었을 때 $O\log(N)$ 시간 복잡도로 업데이트가 가능하다. 샘플링을 진행할 경우에도 동일한 시간 복잡도를 가지기 때문에 학습 데이터 관리에 효과적이고 학습 속도를 개선하는 데 기여할 수 있다. 배치 데이터의 우선순위 p_i는 QPLEX의 TD loss $L(\theta)$기반으로 식(9)에 따라 계산되며 해당 값을 참조해 배치 데이터의 segment tree 위치를 변경한다.

$$p_i = ((L(\theta) \bullet w_i) + \epsilon)^\alpha \tag{9}$$

4.2 update network

본 논문에서는 학습 정책(θ^-)과 행동 정책(θ)이 서로 다른 off-policy 방식으로 모델 학습을 진행한다. 행동 정책은 시나리오 진행 중 실제 행동을 선택하는 정책이고 에이전트는 이를 통해 수집한 데이터를 활용하여 학습 정책을 업데이트하고 개선한다. 또한, 지정된 학습 횟수가 도달할 시점에 행동 정책의 파라미터를 학습 정책에 복사하여 에이전트가 지속적인 성능 향상을 할 수 있게 하였다. 이러한 정책 업데이트 과정에서 손실함수는 에이전트의 학습 효율성을 높이는 중요한 역할을 한다. 1 step TD loss는 에이전트가 현재의 가치 추정치를 업데이트하는 데 사용되며, Dec-POMDP를 반영하여 식(10)과 같이 정의할 수 있다.

$$\begin{aligned} L(\theta) &= E[(\boldsymbol{r} + \gamma V(\boldsymbol{\tau}';\theta^-) - Q(\boldsymbol{\tau}, \boldsymbol{a};\theta))^2] \\ V(\boldsymbol{\tau}';\theta^-) &= \max_{a'} Q(\boldsymbol{\tau}', \boldsymbol{a}';\theta^-) \end{aligned} \tag{10}$$

$\boldsymbol{r}$은 t시점 공동 행동-관찰 이력 $\boldsymbol{\tau}$에서 행동 $\boldsymbol{a}$를 취했을 때의 보상이며, $V(\boldsymbol{\tau}';\theta^-)$는 $t+1$시점 학습 정책이 $\boldsymbol{\tau}'$ 기준으로 가장 높은 행동 가치를 출력하는 행동 $\boldsymbol{a}'$을 선택했을 경우 가치 함수값을 의미한다. 행동 정책의 손실 $L(\theta)$는 배치 데이터의 각 스탭 데이터의 손실 기대값의

평균으로 계산되며 이를 최소화하는 것을 목표로 역전파를 이용해 학습을 진행한다. 이와 같은 모델 학습은 과거 경험에 기반한 정보와 미래의 예상 보상을 균형 있게 반영하여, 정책 성능을 향상시켜 에이전트가 최적의 행동 정책을 찾는데 기여할 수 있다.

IV. Experiment

본 장에서는 III 장에서 정의된 시나리오를 토대로 군집 드론 모델에 MARL 적용 가능성을 검증한다. 또한, 기존 MARL 알고리즘 대비 임무 성공률, 학습 수렴 속도등의 실험 결과를 바탕으로 제안된 알고리즘의 학습 성능 및 학습 효율성을 평가한다.

1. Experiment environment

사전 정의된 시나리오 규칙에 따라 아군, 적군의 드론 수에 차등을 두어 D5_D6, D10_D13으로 시나리오를 구성하였다. 시나리오는 적군 드론 또는 아군 드론이 전부 파괴되거나 사전 정의된 최대 스텝에 도달할 경우 자동 종료되게 설정하였다. 이러한 조건을 바탕으로 아군 드론은 매 스텝마다 자신의 행동에 따른 보상을 받는 방식으로 학습 데이터를 생성한다. 이 보상은 오직 적군 드론의 체력 감소량에 기반하여 계산되며. 이로 인해 누적 보상의 최대치는 고정된다. 이러한 보상체계는 시나리오에서 적군 드론을 모두 파괴해야만 최대 보상을 받을 수 있는 구조로 적군 드론의 섬멸이 승리 조건으로 설정된 상황에 적합하다.

Table 2. proposed algorithm learning parameter

Items		D5_D6	D10_D13
agent network		[90x64x12]	[186x64x19]
hyper network		[131x64x5]	[338x64x13]
MHA	key	[131x1]	[338x1]
	agent	[131x5]	[338x10]
	action	[143x5]	[350x10]
Items		**PER**	
buffer size		4000 scenario	
parameter (α)		0.1	
parameter (β)		0.1	

본 논문에서 제안한 알고리즘 상세 설정은 Table 2를 따르며 정확한 성능 비교를 위해 추가로 기존 MARL 알고리즘 MAPPO, QMIX, QPLEX를 사용하여 군집 드론 모델 학습을 진행하였다. 이때, 학습 알고리즘의 하이퍼 파라미터는 Table 3를 사용하였다. 학습 초반 에이전트는 새로운 시도를 많이 하여 다양한 경험을 쌓게하고 후반에 수렴할 수 있도록 epsilon을 통해 exploration과 exploitation 비율을 조정하였다. 전체 학습 step에서 epsilon rate만큼의 스텝까지 epsilon은 1부터 1차 함수 형태로 감소 되며, 이후 스텝은 minimum epsilon을 사용한다.

Table 3. common hyper-parameter

Items	D5_D6	D10_D13
optimizer	"Adam"	
learning rate (lr)	$5 \cdot 10^{-4}$	
discount factor (γ)	0.99	
minimum epsilon (ϵ)	0.05	
epsilon rate	0.15 %	
policy update cycle	200 epoch	
evaluate cycle	5000 step	
buffer size	4000 scenario	
train batch size	32 scenario	
train epoch	2 epoch	
scenario max step	200 step	
learning max step	3M step	5M step
observation space	90 dim	186 dim
action space	12 dim	19 dim
state space	131 dim	338 dim

2. Experiment result

강화학습은 에이전트가 환경과 상호작용하며 누적 보상을 극대화하는 방향으로 학습을 진행한다. 이 과정에서 에이전트는 각 행동의 결과로 얻는 보상을 기반으로 정책을 업데이트하고, 최적의 행동을 선택하기 위해 경험을 쌓는다. 따라서 군집 드론 시나리오 내 누적 보상과 임무 성공률을 비교 분석하여 제안된 알고리즘의 성능을 평가할 수 있다. 또한 TD loss 변화를 분석하여 학습 과정의 안정성과 수렴 속도를 검증함으로써 알고리즘의 효과를 입증한다. Fig. 6은 제안하는 알고리즘과 기존 MARL 알고리즘의 학습 결과를 보여준다.

군집 드론 모델 학습을 진행한 알고리즘 중 MAPPO를 제외하면 D5_D6, D10_D13 시나리오에서 모두 80% 이상 승률을 기록했다. MAPPO의 경우 학습 초반 다른 알고리즘보다 빠르게 누적 보상 합과 승률이 상승하였지만, 1M step 이후 상승 폭이 낮아지는 것을 볼 수 있다. 이는 에이전트가 학습이 어느 정도 진행되었을 시점에 새로운 시도를 피하는 현상 때문이라고 보인다. 탐험률이 낮아짐에 따라 에이전트는 기존에 학습한 정책을 고수하게 되고 이후 시간이 지나도 결과는 느리게 반영된다. QMIX는 학습 수렴 속도는 느리지만 지속적으로 승률이 증가하는 모습을 볼 수 있고 QPLEX의 경우 안정적으로 학습하는 것을 확인할 수 있다. 제안한 알고리즘은 D5_D6 시나리오에서

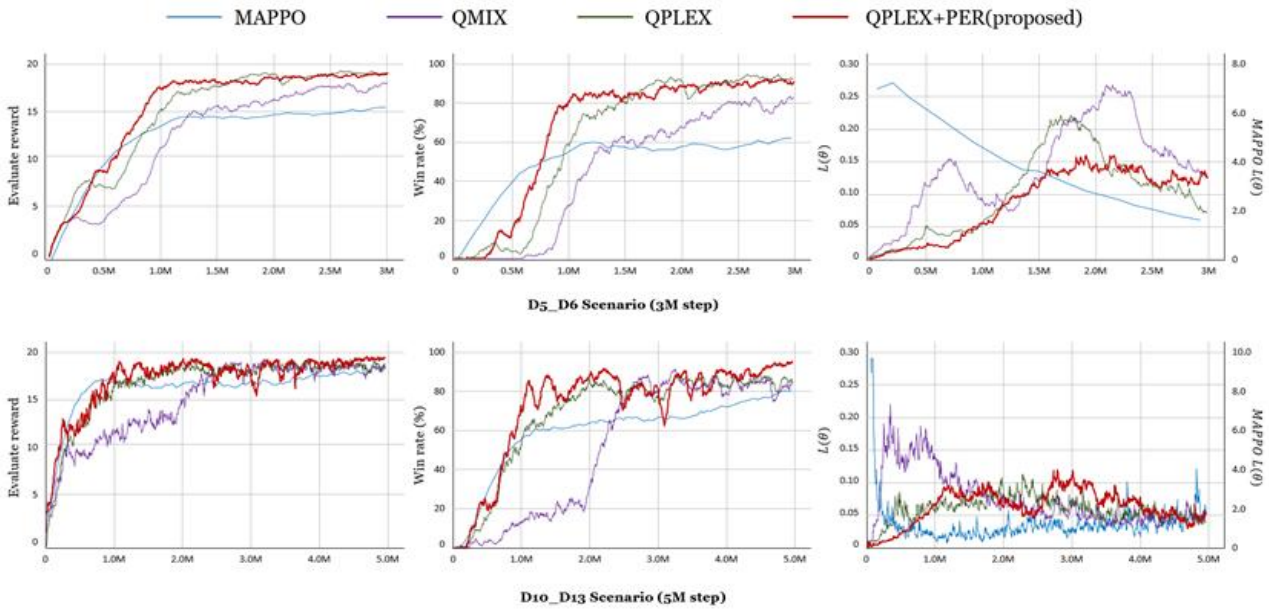

Fig. 6. experiment result

기존 MARL 알고리즘보다 빠르게 수렴하는 것을 확인할 수 있고 D10_D13 시나리오에서는 승률 90%를 초과해 기존 MARL 알고리즘 대비 높은 임무 성공률을 달성하는 것을 확인할 수 있다. 이는 시나리오별 TD loss 기반 가중치를 부여해 데이터의 질에 따라 학습 우선순위를 정하는 PER 기법이 학습 효율성을 높이는 데 기여한 것으로 볼 수 있다. 에이전트의 목표인 학습 정책(θ^{-})과 실제 행동을 선택하는 행동 정책(θ)의 차이가 클수록 데이터 샘플링 우선순위를 높여 지속적으로 학습하기 때문에 보다 가치있는 데이터를 학습하고 그에 따라 학습 수렴 속도가 기존 MARL 알고리즘보다 빠른 것으로 보인다.

제안한 알고리즘은 D10_D13 시나리오 학습 중반 2.5M step에서 loss가 다시 증가하는 모습을 보일 때 누적 보상과 승률도 급격하게 떨어지는 것을 볼 수 있다. 이는 에이전트의 loss가 큰 새로운 경험 데이터를 학습하면서 생기는 현상으로 볼 수 있다. loss가 커지면서 그에 따른 학습 성능도 떨어지는 것처럼 보이지만 학습 가치가 높은 데이터를 학습함으로써 학습 후반으로 갈수록 기존 MARL 알고리즘보다 더 높은 성능을 끌어내는 것을 확인할 수 있다.

Fig.7은 D10_D13 시나리오를 5M step 학습 후 군집 드론 모델의 추론 결과 중 일부이며 이를 통해 군집 드론 모델의 협업 능력을 검증할 수 있다. t_1시점은 적군 드론

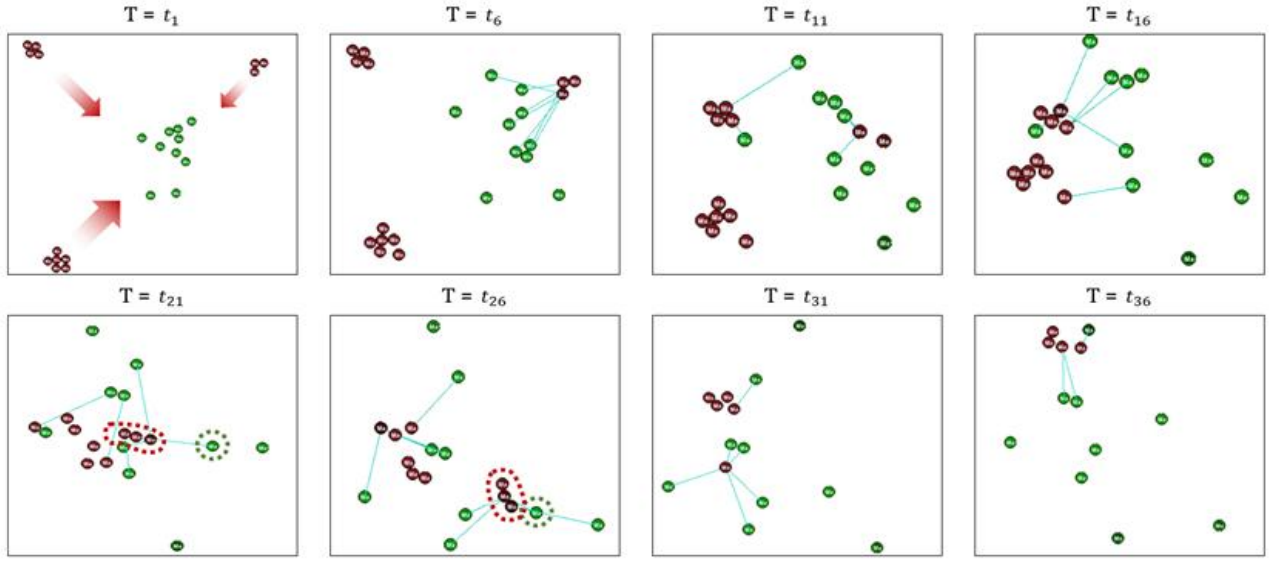

Fig. 7. example of inference in swarm drone model

이 포위된 아군 드론을 향해 3방향으로 공격하는 상태이다. 군집 드론 모델은 적군 드론의 가장 적은 단위 부대를 먼저 식별하고 해당 부대를 먼저 공격하는 모습을 보여준다. 또한, 동일 개체를 집중적으로 공격함으로써(t_6) 빠르게 파괴하고 단위가 큰 부대를 유인해(t_{21}, t_{26}) 전장 상황을 유리하게 만들어 전투를 수행하는 모습을 볼 수 있다.

이와 같은 실험 결과들은 군집 드론 모델에 MARL 적용 가능성을 입증하며, 제안된 알고리즘이 기존 MARL 알고리즘보다 더 높은 임무 성공률과 학습 효율을 보여줌을 확인할 수 있게 해준다. 이러한 결과는 향후 군집 드론의 실제 운영에서 더욱 효과적인 전략 수립에 기여할 것으로 기대된다. 따라서, 제안된 알고리즘은 군집 드론 협업 임무 수행에 있어 실질적인 개선을 가져올 수 있는 가능성을 보여준다.

V. Conclusions

본 논문은 군집 드론 모델의 MARL 알고리즘 적용 가능성을 확인하고 에이전트의 임무 성공률을 높이기 위한 알고리즘을 제안하였다. 학습 네트워크는 QPLEX를 군집 드론 모델에 맞게 수정하였고 데이터별 가중치를 부여하는 PER를 적용하여 학습 데이터 효율을 높였다. 실험에 사용된 교전 시뮬레이터는 SMACv2엔진을 활용하였으며 알고리즘의 정확한 성능 검증을 위해 아군 드론에게 절대적으로 불리한 전장 상황을 전제로 시나리오를 설계하였다. D5_D6 시나리오에서 QPLEX는 1.5M 스텝에서 승률 80%를 달성했고 제안된 알고리즘은 1.0M 스텝에 해당 승률을 달성하여 학습 속도의 증가를 이뤘다. 또한, D10_D13 시나리오에서 제안된 알고리즘으로 학습한 군집 드론 모델이 QPLEX 대비 10% 높은 임무 성공률을 보임으로써 제안된 알고리즘의 성능을 확인할 수 있었다.

향후 연구에서는 제안한 알고리즘의 결과를 바탕으로 모델 경량화 연구를 진행하고자 한다. 드론은 임베디드 보드가 탑재되어 제한된 리소스를 활용하기 때문에 연산량을 줄이기 위한 경량화 연구가 필요하다. 이에 모델 경량화를 통해 군집 드론 모델의 실시간 운영 가능성을 확보하고, 실제 전장 상황에서 효율적으로 작동할 수 있게 개선하고자 한다.

REFERENCES

[1] Seong hyun Yoo, Chun ki Ahn, Jung hun Kim, “Technology and Development of Drones,” The Korea Institute of Electrical Engineers, Vol. 66, No. 2, pp. 19-23, Feb 2017

[2] Seung ho Cheon, Kyung soo Kim, “Defense & Technology”, Korea Defense Industry association, pp. 140-153, Jaunuary 2023.

[3] Tae Joon Park, Yerim Chung, “Multi Objective Vehicle and Drone Routing Problem with Time Window”, Journal of The Korea Society of Computer and Information Vol. 24 No. 1, pp. 167-178, January 2019

[4] Chang-Hun Jiw, Youn-Hee Han, Sung-Tae Moon “Real-Time Hierarchical Deep Reinforcement Learning-Based Drone Trajectory Generation Algorithm Parameter Control”, The Journal of Korean Institute of Communications and Information Sciences '23-10 Vol.48 No.10

[5] Seung won Do, Sung woo Jun, Jae hwan Kim, “Hierarchical Structure Multi-agent Reinforcment Learning for Presenting Battlefield Strategy and Tactics”, IEIE, Summer Annual conference of IEIE, 3,393-3,395, Jeju, korea, June, 2024.

[6] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, Chongjie Zhang, “QPLEX: Duplex Dueling Multi-Agent Q-Learning”, In International Conference on Learning Representations, 2021. DOI: arXiv:2008.01062v3

[7] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. “Prioritized experience replay.”,In ICLR, 2016. DOI: arXiv:1511.05952v4

[8] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, Shimon Whiteson, “The StarCraft Multi-Agent Challenge”, December 2019, DOI: arXiv:1902.04043v5

[9] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, YI WU, “The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games”, 35th Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks DOI: arXiv:2103.01955

[10] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, Thore Graepel, “Value-Decomposition Networks For Cooperative Multi-Agent Learning”, Jun 2017, DOI: arXiv:1706.05296v1

[11] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, Shimon Whiteson, “QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning”, Jun 2018, DOI: arXiv:1803.11485v2

[12] Tonghan Wang, Jianhao Wang, Chongyi Zheng, Chongjie Zhang, “Learning Nearly Decomposable Value Functions Via Communication Minimization”, Jul 2020, DOI: arXiv:1910.05366v2

[13] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel

Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, Shimon Whiteson, “SMACv2: An Improved Benchmark for Cooperative Multi-Agent Reinforcement Learning”, Oct 2023, DOI: arXiv:2212.07489

[14] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, Shimon Whiteson, “SMACv2: An Improved Benchmark for Cooperative Multi-Agent Reinforcement Learning”, Dec 2013, DOI: arXiv:1312.5602v1

Authors

Jin-Ho Ahn received the B.S. degree in information and control engineering from robotics school, Kwangwoon University, Seoul, South Korea, in 2013, and the M.S. degree in control and instrumentation engineering from robotics school in Kwangwoon University, Seoul, South Korea, in 2015. He is currently a Junior Researcher with the Intelligence S/W Team, Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include Muti-Agent reinforcement learnings.

Byung-In Choi received the B.S., M.S., and Ph.D. degrees in electronic engineering from Hanyang University in Seoul, South Korea, in 2001, 2003, and 2008, respectively. He is currently a Leader of the Intelligent S/W team Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include object detection, object tracking, and super resolution.

Tae-Young Lee received the B.S. degree in information and control engineering from robotics school, Kwangwoon University, Seoul, South Korea, in 2009, and the M.S. degree in control and instrumentation engineering from robotics school in Kwangwoon University, Seoul, South Korea, in 2011. He is currently a Senior Researcher with the Intelligent S/W Team, Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include object detection, object tracking and segmentation with deep Learning also generative AI.

Hae-Moon Kim received the B.S. degree in electrical, electronic, and control engineering from Hankyong National University, Anseong, South Korea, in 2020, and the M.S. degree from Computer Science and Engineering, Hanyang University, Ansan, South Korea in 2022. He is currently a Junior Researcher with the Intelligence S/W Team, Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include object detection, instance segmentation, and aerial image processing.

Hyun-Hak Kim received the B.S. and M.S. degrees in biomechanical engineering from Sungkyunkwan University (SKKU), South Korea, in 2020 and 2023, respectively. He is currently a Junior Researcher with the Intelligence S/W Team, Hanwha Systems Co., Ltd., Seongnam, South Korea. His current research interests include reinforcement learnings, object detections, image processing, and language model.