JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 1, pp. 261-270, January 2026
https://doi.org/10.9708/jksci.2026.31.01.261

Reinforcement Learning-Based Multi-Target Threat Assessment and Weapon Assignment Algorithm Using Attention and Responsibility-Based Rewards

Woo-Hyeon Moon*, Seo-Ho Lee*, Won-Seok Jang*, Ji-Seok Yoon*, Hyeon-Mo Kim*, Ju-Mi Park*, Jae-Bok Sung**

*Engineer, InfraTechnology R&D Center, Hanwha Systems, Seongnam, Korea
**Engineer, Reinforcement Learning Team, AgileSoDA, Seoul, Korea

[Abstract]

In this paper, we propose a reinforcement learning-based decision-making algorithm tailored for multi-weapon and multi-threat combat environments. The proposed method separates actor-critic structures by weapon type and incorporates self-attention and cross-attention mechanisms to enable interaction-aware learning across agents. A responsibility-based reward function is designed to evaluate the contribution of each weapon to combat outcomes, promoting cooperative behavior and preventing policy bias. The algorithm is implemented using the Stable-Baselines3 library and validated through tactical simulations. Experimental results demonstrate that our method achieves superior strategic efficiency and collaboration performance compared to conventional PPO-based models, especially under sparse reward conditions and resource constraints.

▸ **Key words**: Reinforcement Learning, Multi-Agent, Threat Assessment, Weapon Allocation, Attention Mechanism, Simulation

[요 약]

본 논문에서는 다중 무장-다중 위협 상황에서의 전략적 판단과 협업적 행동 선택을 위한 강화학습 기반 알고리즘을 제안한다. 제안된 알고리즘은 무장별 actor-critic 구조를 분리하고, self-attention 및 cross-attention 메커니즘을 통해 무장 간 상호작용을 학습한다. 또한, 각 무장의 행동이 전투 결과에 미치는 영향을 정량적으로 평가하는 책임 기반 보상 함수를 설계하여 무장 간 협업을 유도하고, 편향된 정책 학습을 방지한다. Stable-Baselines3 라이브러리를 활용하여 알고리즘을 구현하고, 시뮬레이션 환경에서 기존 PPO 기반 모델과 비교 실험을 수행하였다. 실험 결과, 제안된 알고리즘은 희소 보상 조건과 자원 제약 상황에서도 전략적 효율성과 협업 성능 측면에서 우수한 성능을 보였으며, 실제 전장 시나리오에 적용 가능한 가능성을 확인하였다.

▸ **주제어**: 강화학습, 멀티에이전트, 위협 평가, 무장 할당, 어텐션 메커니즘, 시뮬레이션

• First Author: Woo-Hyeon Moon, Corresponding Author: Jae-Bok Sung
*Woo-Hyeon Moon (moonstar@hanwha.com), InfraTechnology R&D Center, Hanwha Systems
*Seo-Ho Lee (lshpilot@hanwha.com), InfraTechnology R&D Center, Hanwha Systems
*Won-Seok Jang (cws0714g@gmail.com), InfraTechnology R&D Center, Hanwha Systems
*Ji-Seok Yoon (jsyoon2118@hanwha.com), InfraTechnology R&D Center, Hanwha Systems
*Hyeon-Mo Kim (hyeonmo1227@hanwha.com), InfraTechnology R&D Center, Hanwha Systems
*Ju-Mi Park (jumipark1126@hanwha.com), InfraTechnology R&D Center, Hanwha Systems
**Jae-Bok Sung (jaebbok@gmail.com), Reinforcement Learning Team, AgileSoDA
• Received: 2025. 11. 10, Revised: 2025. 12. 16, Accepted: 2025. 12. 23.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

현대 전장 환경은 점점 더 복잡해지고 있으며, 함정, 전투기, 드론, 무인기 등 유무인 복합 체계와 같이 다양한 위협 요소가 동시에 존재하는 다중 위협(multi-threat) 상황에서의 판단과 대응 능력이 전술적 우위를 결정짓는 핵심 요소로 부상하고 있다. 특히 무인 전투 시스템, 자율 무장 플랫폼, 시뮬레이션 기반 전술 훈련 시스템 등에서 인공지능 기반의 상황 인식 및 행동 선택 기술은 필수적인 요소로 자리 잡고 있다[1][2]. 이러한 환경에서 단일 위협(single-threat)만을 고려한 기존의 강화학습 기반 접근 방식으로는 실제 전장 시나리오를 충분히 반영하기 어렵다는 한계가 존재한다.

기존 연구들은 대부분 단일 에이전트가 단일 목표를 추적하거나 단일 위협을 제거하는 상황을 가정하고 있으며[3][4], 복수의 무장 시스템이 동시에 다양한 위협을 평가하고 대응하는 구조에 대한 연구는 상대적으로 부족하다. 예를 들어, [3]에서는 단일 UAV가 고정된 목표를 추적하는 강화학습 기반 경로 최적화 문제를 다루었고, [4]에서는 단일 위협에 대한 전술적 대응을 위한 PPO 기반 정책 학습을 제안하였다. 그러나 실제 전술 환경에서는 함포, 미사일, CIWS 등 다수의 무장 시스템이 존재하며, 이들은 각기 다른 성능과 특성을 가지므로, 단일 위협에 대한 일률적인 대응은 비효율적일 수밖에 없다.

또한, 기존의 보상 함수 설계는 대부분 단순한 생존 시간, 목표 도달 여부, 또는 피해량 감소와 같은 단일 지표에 기반하고 있어, 다중 무장-다중 위협 상황에서의 책임 분배(credit assignment) 문제를 적절히 해결하지 못한다. 이는 무장 간 협업을 유도하지 못하고, 특정 무장만이 반복적으로 선택되는 편향된 정책을 유도할 수 있다. 이러한 한계를 극복하기 위해서는 각 무장의 행동이 전체 전투 결과에 미치는 영향을 정량적으로 평가하고, 이에 기반한 책임 기반 보상 설계가 필요하다.

본 연구에서는 다중 무장-다중 위협 상황에서의 판단 및 행동 선택을 위한 강화학습 기반 알고리즘을 제안한다. 제안하는 알고리즘은 다음과 같은 세 가지 핵심 요소로 구성된다. 첫째, 무장별로 actor-critic 구조를 분리하여 각 무장의 독립적인 정책 학습을 가능하게 하였다. 둘째, self-attention 및 cross-attention 메커니즘을 통해 단일 무장에 대한 정보를 넘어, 다수의 무장 간 상호작용을 학습할 수 있도록 한다. 셋째, 각 무장의 행동이 위협에 미치는 영향과 위협이 아군에 가한 피해를 기반으로 한 책임 기반 보상 함수를 설계하여, 무장 간 협업을 유도하고 정책의 안정성을 향상시킨다.

본 논문의 주요 목적 및 기여점은 다음과 같다:

1) 다중 무장-다중 위협 상황에서의 전략적 판단 및 행동 선택 능력을 강화하기 위해, 무장별 actor-critic 분리 구조와 attention 기반 상호작용 학습을 포함한 강화학습 알고리즘을 제안한다.

2) 책임 기반 보상 설계를 통해 무장 간 협업을 유도하고, 희소 보상 환경에서도 효과적인 정책 학습이 가능함을 실험적으로 검증한다.

3) 강화학습 라이브러리 Stable-Baselines3를 활용하여, 제안한 구조가 실제 시뮬레이션 환경에서 구현 가능하며, 기존 PPO 기반 모델 대비 전략적 효율성과 협업 성능이 향상됨을 확인한다.

본 연구의 목적은 다중 무장-다중 위협 상황에서의 최적 판단 및 행동 선택 능력을 강화하고, 기존 일반적인 강화학습 모델 대비 전략적 효율성과 협업 능력을 향상시키는 것이다. 이를 위해 본 논문은 다음과 같은 순서로 구성된다. 2장에서는 본 연구의 이론적 배경과 관련 연구를 정리하고, 3장에서는 제안하는 네트워크 구조 및 보상 함수 설계를 상세히 설명한다. 4장에서는 시뮬레이션 기반 실험을 통해 제안 모델의 성능을 기존 기법과 비교하고, 5장에서 결론 및 향후 연구 방향을 제시한다.

II. Preliminaries

1. Threat Assessment

위협평가는 전장에 존재하는 다양한 위협 요소들에 대해 그 심각도와 우선순위를 정량적으로 판단하는 과정으로, 전술적 의사결정과 자원 배분의 기준을 제공하는 핵심 기능이다. 특히 다중 위협이 동시에 발생하는 현대 전장 환경에서는 각 위협이 아군에게 미치는 잠재적 피해를 정확히 예측하고, 이에 따라 대응 전략을 수립하는 것이 전투 효율성에 결정적인 영향을 미친다.

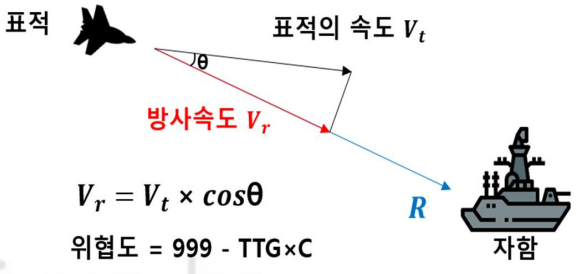

Fig. 1. Target Threat Assessment Using Time To Goal(TTG)

기존의 위협평가 방식은 전문가의 규칙 기반 판단이나 정적 모델링에 의존하는 경우가 많았으며, 이는 복잡하고 동적인 전장 상황에서는 적응성과 확장성이 부족하다는 한계를 가진다. 또한 대부분의 기존 연구는 단일 에이전트가 집중적으로 판단을 하여, 다중 위협 간의 상호작용이나 전술적 연계성을 충분히 반영하지 못한다.

최근 연구에서는 강화학습 기반의 위협평가 모델[2]이 주목받고 있으며, 특히 QMIX 등 협력적 다중 에이전트 강화학습을 통해 희소한 보상 환경에서도 효과적인 위협 인식을 수행하는 모델[5]과 같이 다중 에이전트 환경에서의 협업적 판단 구조가 강조되고 있다.

2. Weapon Assignment

무장할당은 위협평가 결과를 기반으로 각 무장 시스템을 적절한 위협 객체에 배치하는 과정으로, 제한된 자원을 전략적으로 활용하여 전투 효과를 극대화하는 데 목적이 있다. 이는 단순한 목표 추적이나 일률적인 공격이 아닌, 무장 간 협업과 역할 분담을 통해 복잡한 전장 상황에 능동적으로 대응하는 핵심 기능이다.

기존의 무장할당 방식은 휴리스틱 기반의 고정 규칙이나 단일 목표 중심의 최적화 알고리즘에 의존하는 경우가 많았으며, 이는 다중 위협 상황에서의 유연한 대응을 어렵게 만들고, 무장 간 협업을 유도하지 못하는 구조적 한계를 가진다. 특히 단일 에이전트 중심의 강화학습 모델은 복수의 무장 시스템이 존재하는 실제 전장 시나리오를 충분히 반영하지 못하며, 특정 무장에 편향된 정책이 학습되는 문제가 발생할 수 있다.

이에 따라 최근에는 다중 에이전트 강화학습(MARL)을 기반으로 한 무장할당 구조가 제안되고 있으며, PettingZoo 시뮬레이터 기반의 전장 게임에서 연합 강화학습(Federated RL)을 적용하여 무장 간 일반화 성능을 향상하는 연구[6] 등 다양한 전장 지형과 구조에 적응 가능한 분산 학습 방식이 주목받고 있다.

3. Simulator

다중 무장-다중 위협 시뮬레이터는 복수의 무장(에이전트)이 동시에 복수의 위협(적 대상)을 인식하고, 이에 대해 전략적으로 대응하는 시스템을 의미한다. 기존 시뮬레이터는 단일 객체가 단일 적을 대상으로 반복적인 행동을 수행하는 구조가 대부분이었으며, 이는 제한적인 유한 상태 기계(Finite State Machine)를 기반으로 설계되었다[1]. 그러나 실제 전술 환경에서는 다양한 위협이 동시에 등장하며, 각 무장 시스템은 서로 다른 성능과 역할을 가지므로 단일 위협 기반 시뮬레이터는 현실성이 결여된다.

다중 위협 상황에서는 각 무장의 행동이 전체 전투 결과에 미치는 영향이 상호작용적으로 복잡하게 얽혀 있으며, 이를 효과적으로 처리하기 위해서는 무장 간 협업, 위협 간 우선순위 판단 등 다양한 요소가 고려되어야 한다. 특히 무장별로 독립적인 판단을 하면서도 전체 전략적 목표를 공유하는 구조가 요구되며, 이는 기존 제한적인 상태 기반 설계로는 구현이 어렵다. 따라서 본 연구에서는 강화학습 기반의 접근을 통해 이러한 복잡한 상황을 학습 가능한 구조로 모델링하고자 한다.

4. Reinforcement Learning

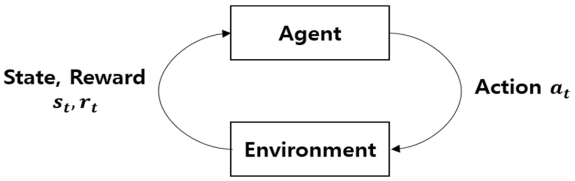

Fig. 2. Training Process of Reinforcement Learning Algorithms

강화학습(Reinforcement Learning, RL)은 에이전트(agent)가 환경(environment)과의 상호작용을 통해 보상(reward)을 최대화하는 방향으로 최적의 행동정책(policy)을 학습하는 기계학습의 한 분야이다.[1][7] 에이전트는 매 시점에서 환경의 상태(state)를 관찰하고, 가능한 행동(action) 중 하나를 선택하여 수행하며, 그 결과로 환경으로부터 보상과 다음 상태를 받는다. 이러한 반복적인 상호작용을 통해 에이전트는 장기적인 누적 보상을 극대화하는 정책을 학습하게 된다.

강화학습은 전통적인 지도학습(supervised learning)과 달리 정답 레이블이 주어지지 않으며, 에이전트가 스스로 시행착오를 통해 최적의 전략을 발견해야 한다는 점에서 차별화된다[1]. 특히 전장 시뮬레이션과 같이 상태 공간이 크고, 행동의 결과가 지연되어 나타나는 환경에서는 강화학습의 순차적 의사결정 능력이 매우 유용하게 작용한다[8].

기존의 강화학습 연구는 주로 단일 에이전트가 단일 목표를 추적하거나 단일 위협을 제거하는 정적 시나리오에 집중되어 왔다. 그러나 실제 전장 환경은 다수의 무장 시스템이 존재하며, 이들은 서로 다른 성능과 제약 조건을 가지는 동시에, 다양한 위협에 대해 협업적으로 대응해야 하는 복잡한 구조를 가진다. 이러한 다중 에이전트-다중위협 상황에서는 단일 에이전트 기반의 강화학습 접근 방식으로는 충분한 전략적 대응이 어렵다[9].

이러한 상황을 대비하기 위해 멀티 에이전트 강화학습(Multi-Agent Reinforcement Learning, MARL) 알고리즘이 나오게 되었다. 멀티 에이전트 강화학습은 복수의 에이전트가 동일한 환경 내에서 상호작용하며, 각자의 정책을 학습하거나 공동의 목표를 달성하기 위해 협력 또는 경쟁하는 강화학습의 확장된 형태이다. 각 에이전트는 독립적으로 상태를 관찰하고 행동을 선택하며, 그 결과로 환경으로부터 보상과 다음 상태를 받아 정책을 갱신한다. 이러한 구조는 단일 에이전트 강화학습으로는 처리하기 어려운 복잡한 상호작용과 전략적 협업을 가능하게 한다.[10]

5. Proximal Policy Optimization

Proximal Policy Optimization(PPO)는 정책 경사(Policy Gradient) 기반 강화학습 알고리즘 중 하나로, 안정성과 구현 용이성을 동시에 만족시키는 방식으로 널리 활용되고 있다[11][12][13]. 기존의 정책 경사 방식은 높은 성능을 보일 수 있지만, 학습 과정에서 정책이 급격하게 변화할 경우 성능이 불안정해지고 수렴하지 않는 문제가 발생한다. 이러한 문제를 해결하기 위해 정책을 신뢰구간 내에서 업데이트하는 Trust Region Policy Optimization(TRPO)[14] 알고리즘이 등장하였으나, 2계 미분(Second-Order)를 계산해야하여 계산 복잡도가 높고 구현이 까다롭다는 단점이 있다.

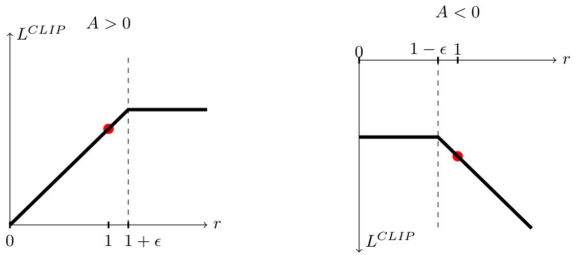

Fig. 3. Illustration of the PPO Clipping Mechanism

PPO는 이러한 문제를 해결하기 위해 clipped surrogate objective를 도입하여, 정책의 변화 폭을 제한하면서도 계산이 간단한 1계 미분(First-Order) 방식으로 구현할 수 있도록 설계되었다. PPO의 핵심 아이디어는 이전 정책과 현재 정책 간의 확률 비율을 계산하고, 이 비율이 일정 범위를 벗어나지 않도록 클리핑(clipping)하여 학습 안정성을 확보하는 것이다. PPO의 기본 손실 함수는 다음과 같이 정의된다:

$$L^{CLIP}(\theta) = \min(r_t(\theta)\widehat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A}_t) \quad (1)$$

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

여기서 $r_t(\theta)$는 $\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$으로 과거 정책에서 새로운 정책으로 업데이트된 비율을 의미하며, $\widehat{A}_t$는 advantage 함수, ϵ은 클리핑 범위이다. 이 수식은 정책이 너무 급격하게 바뀌는 것을 방지하며, 안정적인 학습을 유도한다.

PPO는 클리핑을 통해 정책이 급격히 변하지 않도록 제어함으로써 학습 안정성을 높이는 정책 안정성과 TRPO와 달리 고차 미분이나 Hessian 계산이 필요 없으며, 간단한 gradient descent로 학습이 가능한 계산 효율성의 장점을 가진다. 또한 다양한 환경에서 높은 성능을 보여 일반화 성능이 우수하여 여러 분야에서 사용되고 있다.

III. The Proposed Scheme

본 연구에서는 다중 무장-다중 위협 상황에서의 전략적 판단과 협업적 행동 선택을 가능하게 하기 위해 강화학습 기반의 인공지능 아키텍처를 설계하였다. 제안하는 알고리즘은 크게 인공지능 아키텍처와 상태 및 보상 설계로 구성되며, 각 요소는 전장 시뮬레이션 환경을 반영하여 설계되었다.

1. Artificial Intelligence Architecture

1.1 Separate Hidden Layers

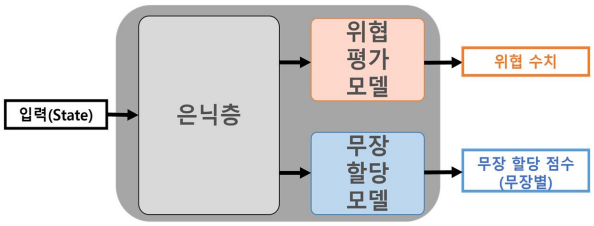

Fig. 4. Separate Hidden Layers Architecture of Threat Assessment and Weapon Assignment Algorithm

본 알고리즘은 위협평가 모델과 무장할당 모델을 서로 연결해주는 공유 은닉층(hidden layer) 구조로 설계하여, 정보 공유를 통해 학습 효율성과 해석 가능성을 높이고자 하였다. 위협평가 모델은 각 위협 객체에 대한 위협 수치를 출력한다. 이 수치는 사용자에게 시각적으로 표시될 수

있으며, 상태 가치 계산에도 활용되어 전략적 판단의 기준으로 작용한다.

무장할당 모델은 각 무장별로 별도의 출력층을 갖는 multi-head 구조로 설계되었다. 이를 통해 각 무장이 개별적으로 할당 대상을 결정할 수 있도록 하였으며, 무장 간의 역할 분담과 협업적 행동 선택을 가능하게 하였다. 이러한 구조는 Transformer의 multi-head 구조를 반영한 것으로, 모든 무장은 각각의 head를 가지고 독립적으로 행동을 결정하도록 설계하였다.

1.2 Attention Mechanism

입력 정보의 표현 학습을 위해 Feature Extract Layer와 Shared Fully Connected Layer를 도입하였다. Feature Extract Layer는 각 개체의 embedding vector에서 유의미한 정보를 추출하며, 입력 형태는 (batch size, ships × embedding dimension), 출력 형태는 (batch size, embedding dimension)으로 설정하였다.

Shared Fully Connected Layer는 위치, 속도 등 고차원의 수치 데이터를 선형 변환을 통해 압축하고 일반화하여, 다양한 입력 벡터 정보를 공통 표현 공간에서 정규화된 형태로 학습할 수 있도록 한다. 입력은 (batch size, ships, information(location, velocity, etc.))형태이며, 출력은 (batch size, ships, embedding dimension)로 구성된다.

이후 self-attention 메커니즘을 통해 개체 간의 관계를 학습하고, 각 무장 head에서는 cross-attention을 적용하여 무장 간의 협업 가능성을 강화하였다. 이를 통해 단순한 개체별 판단을 넘어, 전체 전장 상황에 대한 글로벌 컨텍스트를 반영한 전략적 행동 선택이 가능하도록 하였다.

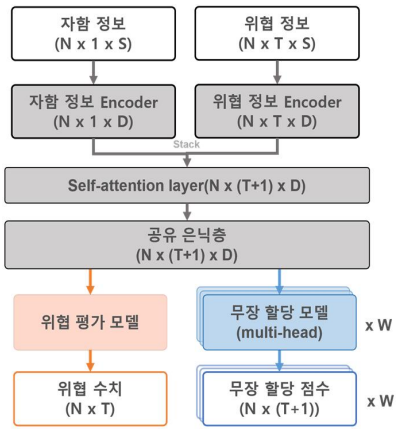

Fig. 5. Multi-Threat Battlefield Threat Assessment and Weapon Assignment Algorithm Architecture

2. State and Reward Design

2.1 State Representation

상태 정보는 자함(unit), 무장(weapon), 위협(threat) 객체 각각에 대해 세분화된 특성으로 구성된다. 자함의 상태는 유닛 타입, 파괴 여부, 상대좌표(x, y), 속도(x, y), 내구도, 장착 무장 수 등으로 구성되며, 각 무장에 대한 weapon state는 무장 타입, 직전 타겟 좌표, 공격 성공 여부, 장전 여부, 사거리, 위력, 장전 진행도, 잔탄수 등 총 8개의 항목으로 정의된다.

전체 상태는 다음과 같이 표현된다:

1) state_ally [1, S]: 자함의 상태로, unit state와 weapon state를 무장 수만큼 이어 붙인 형태(S: 자함 및 적함 state 벡터의 길이)

2) state_threat [T, S]: 적함의 상태로, 자함과 동일한 구조(T: 위협의 수)

3) state_weapon [W, S_w]: 자함의 무장 상태로, weapon state 기반(W: 자함 무장 수, S_w : weapon state 벡터의 길이)

이외에도 행동(action), 마스크(mask), 보상(reward), log_prob, advantage, value_target 등은 학습 과정에서 상대적 크기나 확률 분포, Generalized Advantage Estimate(GAE) 기반 계산을 통해 의미를 갖는다. 평가 단계에서는 미리 생성된 에피소드의 초기 상태를 입력으로 사용하며, 출력은 평균 피해량, shaped reward, 페널티로 구성된다.

2.2 Reward Function

보상 함수는 위협이 자함에 가한 피해를 기반으로 음의 보상을 부과하고, 이를 최소화하는 방향으로 학습이 진행되도록 설계하였다. 피해(D)가 발생한 경우, 직전 step까지 해당 위협을 공격할 수 있었던 무장에 대해 책임을 분배하여 음의 보상을 부여한다. 이를 통해 무장 간의 역할 분담을 유도하고, 특정 무장에 편향된 정책 학습을 방지한다.

	1	2	3	4	5	6	7	8	9	10
step별 획득 보상 #1	0	0	0	0	0	0	0	-D/8	$-D/8+I_1$	$-D/8+I_1$
step별 획득 보상 #2	$-D/8+I_2$	-D/8	-D/8	-D/8	-D/8	0	0	0	0	0

Fig. 6. The Reward Function Architecture of Proposed Idea

또한, 무장이 위협에 가한 피해량을 기반으로 양의 보상(I)을 부여함으로써, 실제 공격을 수행한 경우에는 그 책임을 완화하고 과도하게 방어적인 정책 학습을 방지한다. 이로 인해 학습 효율성과 전략적 다양성이 향상된다.

보상 함수는 다음과 같이 정의된다:

$$reward(w_n, s_m) = \sum_{t_j}^{T}\left(\frac{M(w_n, t_j, s_m)D(t_j, s_m)}{\sum_{w_i}^{W}\sum_{s_k}^{S}M(w_n, t_j, s_m)} + \alpha \times I(w_n, t_j, s_m)\right) \quad (2)$$

$D(t,s) = step\,s$에 위협 t가 자함에 가한 피해

$M(w,t,s) = \begin{cases} 1, & step\,s\text{에 무장 } w\text{가 위협 } t\text{를 공격 가능한 경우} \\ 0, & otherwise \end{cases}$

$I(w,t,s) = step\,s$에 무장 w가 위협 t에 가한 피해

$W =$ 무장 집합

$T =$ 위협 집합

$S =$ 책임 분배 대상 $step$의 집합

$\alpha =$ 양의 보상 계수

여기서 D(t,s)는 위협 t가 step s에서 아군에게 가한 피해(음수)도를, M(w,t,s)는 step s에서 무장 w가 위협 t를 공격 가능한지 여부를(1, 0) 나타내는 마스크 함수를, I(w,t,s)는 step s에서 무장 w가 위협 t에 가한 피해량을 나타낸다.

IV. Experiments

1. Experiment Environment & Scenario

본 연구에서는 제안한 강화학습 기반 다중 무장-다중 위협 판단 및 행동 선택 알고리즘의 전략적 효율성과 협업 능력을 평가하기 위해, 구조화된 해상 전술 시뮬레이션 환경을 설계하였다. 해당 시뮬레이터는 2차원 평면상에서 작동하며, 자함(아군 함선)은 고정된 위치에서 다양한 위협에 대응하는 시나리오 기반의 전투 상황을 모사한다. 각 시나리오는 무장 구성, 위협 특성, 자원 제한 조건 등을 달리하여 에이전트의 전략적 판단 능력을 다각도로 검증할 수 있도록 설계되었다.

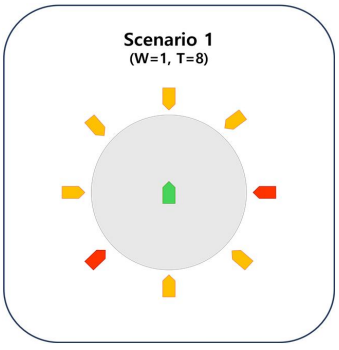

Fig. 7. Scenario 1: Evaluation of Basic Response Capability Using a Single Weapon System

시나리오 1은 강화학습 에이전트가 단일 무장을 활용하여 다방향에서 접근하는 위협에 대응하는 능력을 평가하기 위한 환경이다. 자함은 중앙 좌표 (500, 500)에 고정된 위치에서 시작하며, 이동 능력은 없고 함포를 탑재하고 있다. 해당 무장은 사거리 50, 공격력 4, 정확도 100%의 성능을 가지며, 360도 회전이 가능하여 모든 방향의 위협에 대응할 수 있다. 재장전 속도는 1로 설정되어 있어 빠른 연속 사격이 가능하며, 최대 200발의 탄약을 보유한다. 본 시나리오는 에이전트가 단일 무장을 활용하여 위협의 위치와 접근 속도에 따라 효과적인 대응 전략을 학습할 수 있는지를 평가한다.

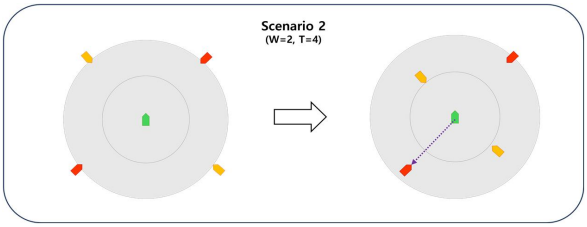

Fig. 8. Scenario 2: Evaluation of Strategic Weapon Selection Under Resource Constraints

시나리오 2는 에이전트가 제한된 자원을 전략적으로 할당하고, 다양한 위협에 효과적으로 대응할 수 있는지를 평가하기 위해 설계되었다. 자함은 함포와 장거리 함포 이렇게 두 종류의 무장을 탑재하고 있으며, 각각 사거리 75, 공격력 1의 기본형 무장과 사거리 120, 공격력 5의 장거리 고성능 무장으로 구성된다. 특히 장거리 함포는 최대 탄약 수량이 4발로 제한되어 있어, 자원 관리 및 전략적 사용이 중요한 요소로 작용한다.

자함을 둘러싼 위협 객체는 총 4대로, 자함과 동일한 거리에서 원형으로 배치되며 자율적으로 이동하며 공격을 시도한다. 이 중 2대는 근거리 무장 함포(사거리 50, 공격력 1)를 탑재하고 있으며, 나머지 2대는 고위험 무장 함포(공격력 10)을 탑재하고 있다. 시뮬레이션은 최대 100 스텝으로 구성되며, 각 스텝마다 에이전트는 관찰, 행동, 보상 과정을 반복한다. 보상은 피해 경감과 공격 기여도를 기반으로 계산되며, 잘못된 행동에 대해서는 -2.0의 패널티가 부여된다. 본 시나리오의 핵심은 에이전트가 제한된 탄약을 고위험 위협에 집중적으로 할당하는 전략을 학습할 수 있는지를 평가하는 데 있다.

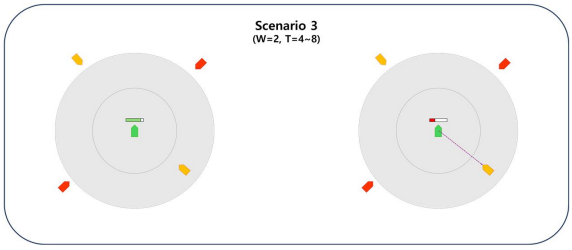

Fig. 9. Scenario 3: Evaluation of Resource Allocation and Survival Strategy in a Complex Threat Environment

시나리오 3은 자함이 다양한 성능의 무장을 탑재한 다수의 위협 객체에 대응해야 하는 복잡한 전술 환경을 모사한다. 자함은 함포(사거리 75, 공격력 2)와 장거리 함포(사거리 120, 공격력 5, 탄약 10발)를 탑재하고 있으며, 내구도는 20~60 사이의 범위로 설정되어 다양한 생존 조건을 실험할 수 있도록 하였다. 자함은 전방위 사격이 가능하며, 모든 방향에서 접근하는 적을 타격할 수 있다.

해당 시뮬레이션에서는 위협 객체 4대가 자함과 동일한 거리에서 배치된다. 이 중 2대는 고위협 함포를 탑재하여 일반 무장 대비 2배 이상의 피해를 가할 수 있으며, 나머지 2대는 단거리 함포를 탑재하여 근거리에서 빠르게 접근해 공격을 시도한다. 각 위협 객체는 고유한 시작 좌표와 속도(10~20)를 가지며, 자율적으로 이동하며 자함을 공격한다. 시뮬레이션은 최대 120 스텝으로 구성되며, 각 스텝마다 에이전트는 관찰, 행동, 보상 과정을 반복한다. 보상은 피해 경감과 공격 기여도를 기반으로 계산되며, 잘못된 행동에 대해서는 -2.0의 패널티가 부여된다.

본 시나리오의 핵심은 에이전트가 위협의 무장 구성과 위치, 이동 속도를 종합적으로 고려하여 전략적으로 무장을 선택하고, 제한된 자원을 고위험 위협에 집중적으로 할당함으로써 전체 전투 효율성과 생존 가능성을 극대화하는 전략을 학습할 수 있는지를 평가하는 데 있다.

2. Experiment Results

본 장에서는 제안한 강화학습 기반 무장할당 알고리즘의 성능을 정량적으로 평가하기 위해 수행한 실험 결과를 기술한다. 실험은 앞서 정의한 세 가지 시나리오를 기반으로 진행되었으며, 제안한 알고리즘(Ours PPO)과 기존 PPO 알고리즘(Traditional PPO)을 비교 분석하였다. 이때 기존 PPO 알고리즘은 Stable-Baselines3 라이브러리를 참고하여 제작하였다. 평가 항목은 총 일곱 가지로 구성되며, 각 지표는 다음과 같은 의미를 가진다.

우선, Actor Loss는 에이전트가 특정 행동을 선택할 확률이 과거 정책 대비 얼마나 변화했는지를 advantage와 결합하여 측정한 값이다. PPO 알고리즘에서는 advantage를 최대화하는 것이 목표이므로, actor loss는 부호를 반전시켜 최소화하는 방향으로 학습이 진행된다. Critic Loss는 GAE(Generalized Advantage Estimation)를 통해 계산된 critic target과 현재 critic network의 출력값 간의 평균제곱오차(MSE)를 의미하며, 가치 함수의 정확도를 평가하는 지표이다. Entropy Loss는 정책의 행동 확률 분포가 지나치게 빠르게 수렴하는 것을 방지하기 위한 항으로, 탐색성을 유지하는 데 기여한다.

Penalty는 잘못된 무장 할당이나 작전 실패 등 전략적 오류에 대해 부과되는 음의 보상이며, Score는 자함이 받은 피해량에 비례하는 수치로, 에피소드 내 총 피해량에 계수를 곱하여 계산된다. Shaped Rewards는 자함이 받은 피해에 따른 음의 보상을 책임 비중에 따라 각 무장에 분배하고, 가한 피해 및 penalty에 계수를 곱해 더한 최종적인 보상으로 정의된다. 마지막으로 Total Loss는 actor loss, critic loss, entropy loss에 각각의 계수를 곱하여 합산한 전체 손실 값으로, 학습 안정성과 수렴 속도를 종합적으로 판단하는 데 활용된다.

실험 결과, Fig. 10.와 Table 1, Table 2과 같이, 제안한 알고리즘은 대부분의 시나리오에서 기존 PPO 기반 알고리즘 대비 우수한 성능을 보였다. 여기서 Fig. 10.은 각 시나리오별 제안한 알고리즘(ours)과 벤치마크 알고리즘(ppo)의 성능을 비교하는 그래프로, x축은 강화학습이 진행된 step을 의미하고, y축은 평가지표를 의미한다. 본 연구에서는 다양한 지표를 활용하여 성능을 평가하였으나, 그 중에서도 Score와 Total Loss를 핵심 성능 지표로 해석하였다. Score는 위 2.2 Reward Function에서 언급하였듯이, 강화학습 에이전트의 보상함수 값을 의미하고, Total Loss는 학습 과정에서의 안정성과 정책 수렴 정도를 보여주는 지표로서, Actor Loss, Critic Loss, 그리고 Entropy Loss를 합한 값이다.

시나리오 1에서는 제안한 알고리즘이 기존 PPO 대비 Score에서 약 10% 향상된 결과를 보였으며, Total Loss는 82% 감소하여 학습 안정성이 크게 개선되었다. Actor loss와 Critic loss가 낮게 유지되었고, penalty 역시 거의 발생하지 않아 불필요한 행동 선택이 억제되었음을 확인할 수 있었다. 이는 제안한 알고리즘이 단일 위협 상황에서도 보다 효율적인 전략을 학습했음을 의미한다.

시나리오 2에서는 자원 제약이 있는 환경에서 전략적 판단 능력을 평가하였다. 이 경우 제안한 알고리즘은 Score에서 기존 PPO 대비 28% 향상된 결과를 기록하였으며, Total Loss는 88% 감소하여 자원 제약 조건에서도 안정적인 정책 학습이 가능함을 보여주었다. 특히 penalty

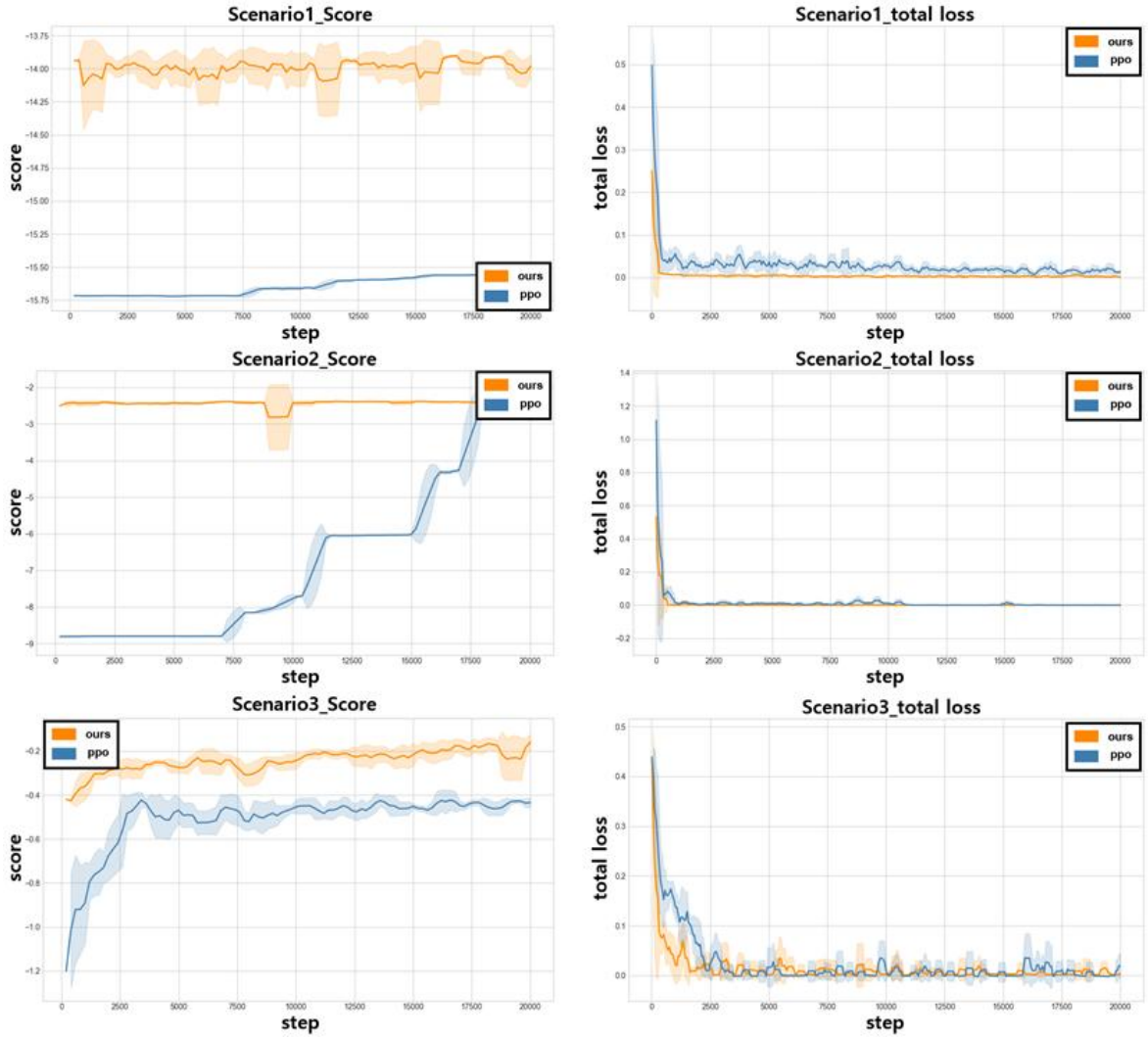

Fig. 10. The Result of Multi-Threat Battlefield Threat Assessment and Weapon Assignment Algorithm Architecture

Table 1. The Result Of Proposed Algorithm(Ours PPO)

Scenario	Actor Loss	Critic Loss	Ent Loss	Penalty	Score	Shape Rewards	Total Loss
1	0.000637	0.010755	0.021794	-0.001538	-13.965115	1.016673	0.002788
2	-0.000151	0.000954	0.010358	0.000000	-2.395769	-0.395769	0.000058
3	0.000686	0.040258	0.015433	-0.324615	-0.190962	2.472019	0.004809

Table 2. The Result Of Traditional Algorithm(PPO)

Scenario	Actor Loss	Critic Loss	Ent Loss	Penalty	Score	Shape Rewards	Total Loss
1	-0.000177	0.079256	0.002012	-0.396154	-15.552404	0.008721	0.015674
2	-0.000412	0.004671	0.001838	-0.023077	-3.335865	-1.001038	0.000521
3	-0.000966	0.038297	0.018827	-0.400000	-0.442192	2.18349	0.007678

가 발생하지 않았다는 점은 제안한 알고리즘이 자원 활용의 효율성을 극대화하면서 불필요한 손실을 최소화했음을 시사한다. Actor loss와 Critic loss 역시 기존 PPO 대비 낮은 수준을 유지하여, 정책과 가치 함수의 학습이 균형적으로 이루어졌음을 확인할 수 있었다.

시나리오 3은 복합 위협 환경에서의 자원 배분과 생존 전략을 평가하기 위한 실험으로, 다양한 위협 유형과 제한된 탄약 조건이 포함되었다. 이 시나리오에서 제안한 알고리즘은 Score에서 기존 PPO 대비 56% 향상된 결과를 기록하였으며, Total Loss는 37% 감소하였다. Critic loss

가 다소 높게 나타났으나, 이는 복합 위협 환경에서 가치 함수가 다양한 상황을 평가해야 하는 특성에 기인한 것으로 해석된다. 그럼에도 불구하고 shaped reward가 기존 대비 높은 수준을 보였다는 점은 제안한 알고리즘이 위협 간 우선순위를 효과적으로 학습하고, 자원 배분 전략을 최적화했음을 보여준다.

종합적으로, 제안한 알고리즘은 단순한 수치 개선을 넘어 학습 과정에서 안정성을 확보하고, 다양한 전술 환경에서 전략적 효율성과 협업 능력을 강화하는 흐름을 나타냈다. 특히 Score와 Total Loss의 개선은 제안한 알고리즘이 실제 전술적 의사결정 상황에서 기존 PPO 대비 더 높은 성능을 발휘할 수 있음을 실험적으로 입증한 결과라 할 수 있다.

V. Conclusions

본 논문에서는 다중 무장-다중 위협 상황에서의 전략적 판단과 협업적 행동 선택을 가능하게 하기 위해, self-attention과 책임 기반 보상 구조를 포함한 강화학습 알고리즘을 제안하였다. 위협평가와 무장할당 모델을 분리하고, multi-head actor 구조와 attention 메커니즘을 통해 무장 간 협업과 역할 분담을 학습할 수 있도록 설계하였다.

실험 결과, 제안한 알고리즘은 기존 PPO 대비 낮은 손실값과 높은 shaped reward를 기록하며 전략적 효율성과 학습 안정성 측면에서 우수한 성능을 보였다. 특히 제한된 자원 조건과 복합 위협 환경에서 에이전트가 고위험 위협에 효과적으로 대응하고, 장거리 무장을 전략적으로 할당하는 행동을 학습한 것을 확인하였다.

향후 연구에서는 실제 전장환경과 유사한 고도화된 시뮬레이터를 제작하고 본 알고리즘을 시뮬레이션에 적용하여 다양한 무장 및 위협 유형에 대한 적응형 정책 학습을 통해 자율 전투 시스템의 실용 가능성을 확대할 예정이다.

REFERENCES

[1] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT press, 1998.

[2] Woohyeon Moon, Wonseok Jang, Haseok Song. "Proposal of the Reinforcement Learning-Based Threat Assessment Algorithm for Anti-Ship and Anti-Aircraft Targets in Naval Combat Management System." Proceedings of the Korea Institute of Military Science and Technology Conference, pp. 1303-1304, Jeju, Republic of Korea, June 2024.

[3] Zhang, Kaiqing, Zhuoran Yang, and Tamer Başar. "Multi-agent reinforcement learning: A selective overview of theories and algorithms." Handbook of reinforcement learning and control (2021): 321-384.

[4] Lowe, Ryan, et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." Advances in neural information processing systems 30 (2017).

[5] Kim, Minkyoung. "Cooperative Multi-agent Reinforcement Learning on Sparse Reward Battlefield Environment using QMIX and RND in Ray RLlib." Journal of The Korea Society of Computer and Information 29.1 (2024): 11-19.

[6] Mozumder, Fardeen Hasib. "Multi-Agent Battlefield Game with Federated Reinforcement Learning." 2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM). IEEE, 2025.

[7] Woo-Hyeon Moon, & Yun-Su Kim (2020-10-22). Comparison and Analysis of Artificial Intelligence Techniques for Optimal Core Shape Design of Wireless Power Transfer System. Proceedings of the KIEE Conference, Jeju, Republic of Korea.

[8] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." nature 518.7540 (2015): 529-533.

[9] Busoniu, Lucian, Robert Babuska, and Bart De Schutter. "A comprehensive survey of multiagent reinforcement learning." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38.2 (2008): 156-172.

[10] Busoniu, Lucian, Robert Babuska, and Bart De Schutter. "A comprehensive survey of multiagent reinforcement learning." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38.2 (2008): 156-172.

[11] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[12] Moon, Woohyeon, et al. "Path planning of cleaning robot with reinforcement learning." 2022 IEEE International Symposium on Robotic and Sensors Environments (ROSE). IEEE, 2022.

[13] Moon, Woohyeon, et al. "Enhanced Transformer Architecture for Natural Language Processing." arXiv preprint arXiv:2310.10930 (2023).

[14] Schulman, John, et al. "Trust region policy optimization." International conference on machine learning. PMLR, 2015.

Authors

Woo-Hyeon Moon received the B.S. degrees in Electronic Engineering from Jeonbuk National University and M.S. degrees in Artificial Intelligence(Reinforcement Learning, Natual Language Processing)-based Robotics from Korea Advanced Institute of Science and Technology, South Korea, in 2022 and 2024. He is currently with Hanwha Systems Co. from 2024. His research interests include Intelligent Combat Management System, Reinforcement Learning, Multi-Agent Reinforcement Learning, Natual Language Processing, Large Language Model, and Artificial Intelligence.

Seo-Ho Lee received the B.S. degrees in Mechanical Engineering and Unmanned Vehicle Engineering from Sejong University, Korea, in 2022, and the M.S. degrees in Intelligent Mechatronics from Sejong University, Korea, in 2024. He is currently with Hanwha Systems Co. from 2024. His research interests include Intelligent Naval Combat Management System, Control Engineering, and Reinforcement Learning.

Won-Seok Jang received M.S degree in Computer Engineering from Chungnam National University, South Korea and completed Ph.D. program from Kyungpook National University, Korea. He is currently working in Hanwha Systems Co. from 2017. He is interested in Intelligent Naval Combat Management System using Artificial Intelligence, Naval Unmanned System and Naval Safety System.

Ji-Seok Yoon received the B.S degree in Electronic Engineering from Kumoh National Institute of Technology, South Korea, in 2012, and the M.S. and Ph.D. degrees in mechatronics from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2014 and 2021. From 2021 to 2023, he was a Postdoctoral Researcher with the Vision and Image Processing Laboratory, Tech University of Korea. Since 2023, he has been the CTO of IKLAB Inc. Currently, he is working in Hanwha Systems Co. from 2023. He is interested in Intelligent Naval Combat Management System using Artificial Intelligence.

Hyeon-Mo Kim received the B.S. and M.S. degrees in Computer Science and Engineering from Sejong University, South Korea, in 2022 and 2024. He is currently with Hanwha Systems Co. from 2024. He is currently with Hanwha Systems Co. from 2024. His research interests include Intelligent Naval Combat Management System, Data Mining, Ontology Engineering, Large Language Model, and Artificial Intelligence.

Ju-Mi Park recieved the B.S degree in Divsion of Media, Culture and Design Technology from Hanyang University ERICA, South Korea and the M.S. degree in School of Integrated Technology from Gwangju Institute of Science and Technology, Korea. She is currently working in Hanwha Systems Co. from 2023. She is interested in Intelligent Naval Combat System using Artificial Intelligence.

Jae-Bok Sung received the B.S. degree in Engineering in 2021 and the M.S. degree in Engineering in 2024 from the University of Electro-Communications, Japan. He is currently a research engineer at AgileSoDA. His research interests include risk-sensitive reinforcement learning and robust decision-making under uncertainty.