https://doi.org/10.7840/kics.2026.51.1.209

AI 화력운용참모: 보상 적응형 강화학습 기반 군사 지휘결심 지원 시스템

이 재 휘•, 엄 찬 인*, 김 찬**, 김 경 수**, 이 형 도**, 강 현 수**, 권 민 혜°

AI Fire Support Officer: Military Decision Support System Based on Reward Adaptive Reinforcement Learning

Jaehwi Lee•, Chanin Eom*, Chan Kim**, Kyeongsoo Kim**,
Hyeongdo Lee**, Hyunsu Kang**, Minhae Kwon°

요 약

최근 군사 의사결정 지원 분야에서는 복잡한 전장 의사결정을 자동화하기 위해 심층 강화학습 기반 접근이 활발히 연구되고 있다. 본 논문에서는 지휘결심 지원을 위한 보상 적응형 강화학습 기반 화력운용 시스템을 제안한다. 제안된 시스템은 전장의 상황을 인지하고, 인지된 정보를 바탕으로 지휘관의 요망효과를 달성하기 위한 최적의 의사결정을 수행하도록 설계하였다. 강화학습 기반의 지휘결심 의사결정 방법으로 사전 수집 데이터와 온라인 상호작용 데이터를 통합적으로 활용하며, 보상 정보를 기반으로 한 선택적 모방 학습을 통해 정책의 효율성과 안정성을 동시에 확보한다. 다양한 전장 시나리오를 모사한 시뮬레이션 환경에서 수행한 실험 결과, 제안된 정책은 기존 강화학습 및 휴리스틱 기반 방법 대비 평균 임무 달성률을 약 29% 향상시켰으며, 주어진 제약조건을 준수하면서도 높은 임무 수행 효율을 달성하는 것을 확인하였다.

키워드 : 군사 의사결정, 지휘결심 의사결정 지원, 무기 - 표적 할당, 강화학습

Key Words : Military decision-making, Command decision support, Weapon-target assignment, Reinforcement learning

ABSTRACT

Recent studies in military decision support have actively explored deep reinforcement learning (RL) approaches to automate complex battlefield decision-making processes. This paper proposes a reward-adaptive RL-based firepower operation system designed to support command decisions in dynamic combat environments. The proposed system perceives battlefield situations through a perception module and makes decisions to achieve the commander's desired effects. The decision-making module integrates both pre-collected and online interaction data while employing a reward-adaptive selective imitation mechanism to enhance sample efficiency and stability simultaneously. Through simulated battlefield scenarios, the proposed system demonstrated an average 29% improvement in mission achievement compared to conventional RL and heuristic-based methods, while effectively satisfying given operational constraints.

◆ First Author : Soongsil University Department of Intelligent Semiconductors, jaehwilee@soongsil.ac.kr, 학생회원
° Corresponding Author : Soongsil University Department of Intelligent Semiconductors and School of Electronic Engineering, minhae@ssu.ac.kr, 종신회원
* Soongsil University Department of Intelligent Semiconductors, chanineom@soongsil.ac.kr, 학생회원
** Konan Technology Inc., chan.kim@konantech.com; kyeongsoo.kim@konantech.com; hyeongdo.lee@konantech.com; hyunsu.kang@konantech.com
논문번호 : 202511-294-0-SE, Received October 31, 2025; Revised November 27, 2025; Accepted Dcember 4, 2025

Ⅰ. 서 론

최근 군사작전 환경은 무인화, 지능화, 네트워크화가 동시에 심화되면서 지휘결심을 지원하는 지능형 의사결정 시스템의 필요성이 크게 증대되고 있다[1-3]. 현대 전장에서 지휘관은 제한된 시간과 자원 내에서 최적의 결정을 내려야 하며, 이를 지원하기 위한 인공지능 기반의 기술은 임무 달성과 다양한 군사적 제약 요소들을 동시에 고려할 수 있는 핵심 요소로 주목받고 있다[4,5].

지휘결심 지원 시스템은 전장 각 요소가 유기적으로 연결된 네트워크 중심전을 기반으로 전장의 복잡한 정보를 실시간으로 수집 및 처리하고, 이를 토대로 최적의 결심을 도출하는 시스템이다[1]. 정찰 드론, 감시 장비, 지상 센서 등 다양한 정보원을 통해 전장 상황을 인지하고, 확보된 데이터는 지상 제어국과의 네트워크 체계를 통해 전달된다. 이 과정을 통해 시스템은 표적의 우선순위, 아군의 가용 자원, 군사적 제약조건 등을 종합적으로 고려하여 지휘관의 의사결정을 보조하는 핵심 기능을 수행한다. 특히, 이러한 복잡한 전장 환경에서 최적의 행동을 산출할 수 있는 지능형 의사결정 모듈은 필수적이다. 이 모듈은 인지 데이터를 입력으로 활용하여 임무 목표를 최대화하는 방향으로 결정을 내리는 역할을 수행해야 한다. 이러한 점에서, 환경과의 상호작용을 통해 최적 정책을 학습하는 강화학습은 지휘결심 지원 시스템의 핵심 기술로 주목받고 있다[4,5].

지휘결심 지원 시스템에서는 지휘관의 요망효과를 달성하는 동시에 임무 수행 과정에서 주어진 제약조건을 만족시키는 것이 핵심 과제로 고려된다. 이를 위해 심층 강화학습이 활용되며, 이러한 접근은 수집된 데이터의 경향을 모방하는 학습 방식과 마르코프 의사결정 과정(Markov Decision Process; MDP) 모델 설계를 통해 구현될 수 있다[6,7]. 학습 개체는 주어진 상태(state)에서 행동(action)을 수행하고 보상 함수(reward function)를 통해 피드백을 받아 정책(policy)을 개선하므로, 해결하고자 하는 문제의 목표에 부합하는 MDP 설계는 매우 중요하다.

그러나 실제 군사작전 환경에서는 실전 데이터를 직접 수집하기 어렵고, 보안 및 기밀 문제로 인해 데이터 접근성이 낮다. 이러한 제약으로 인해 강화학습의 온라인 상호작용 기반 학습을 직접 적용하기 어렵기 때문에, 사전 수집된 데이터셋을 활용한 데이터 기반 강화학습이 효과적인 대안으로 주목받고 있다[8-10]. 이러한 접근은 다양한 정책에서 수집된 데이터를 학습에 반영함으로써 학습 안정성과 데이터 효율성을 동시에 향상시킬 수 있다. 특히 모방학습(Behavioral Cloning; BC) 기반의 초기 정책 학습이나, 강화학습 과정에서 BC 항을 정규화 요소로 결합하는 방식은 데이터 품질에 따른 정책 신뢰성을 높이는 데 효과적이다.

본 연구에서는 데이터 기반 강화학습 방법을 활용하여 지휘결심 지원을 위한 화력운용 의사결정 시스템을 제안한다. 제안하는 시스템의 목표는 전장 환경의 제약조건을 고려하며, 지휘관의 요망효과를 달성하기 위해 탄종과 사용량을 효율적으로 결정하는 것이다. 이를 위해 전장 환경을 MDP로 모델링하고, 사전 수집된 데이터셋을 효율적으로 활용하기 위해 강화학습과 모방학습을 결합한 보상 적응형 학습 방식을 도입한다.

본 논문의 주요 기여는 다음과 같다.

- 지휘관의 요망효과와 군사 제약을 반영할 수 있는 지휘결심 지원 체계 시스템의 화력 조합 결정 문제에 대한 MDP를 설계하였다.
- 강화학습과 모방학습을 결합하여 사전 수집된 데이터셋을 효율적으로 활용할 수 있는 학습 방식을 설계하고, 데이터 품질에 기반한 보상 조정 기법을 적용하였다.
- 실제 군에서 사용되는 표적 처리 절차를 반영한 시뮬레이션 환경을 구현하고, 이를 통해 제안한 시스템의 성능을 평가하였다.
- 제안된 방법이 규칙 기반 모델 및 기존 모델들과 비교하여 지휘관의 성향을 반영한 맞춤형 의사결정과 자원 활용 효율성 측면에서 우수한 성능을 보임을 실험적으로 검증하였다.

본 논문의 구성은 다음과 같다. II장에서 본 연구의 선행연구에 대해 살펴보고, III장에서는 본 연구에서 제안하는 시스템의 구조와 구성요소들에 대해 설명한다. IV장에서는 제안 시스템의 인지 모듈의 학습 방법 및 강화학습 기반 지휘결심을 위한 MDP 모델과 시스템에 적용된 알고리즘을 소개한다. V장에서는 제안 시스템의 인지 모듈에 대한 성능 평가를 진행하며, VI장에서는 강화학습 의사결정 모델의 성능을 평가한다. 마지막으로, VII장에서는 본 연구의 결론을 맺는다. 본 논문에서 사용된 모든 기호와 표기법은 Appendix A에서 확인할 수 있다.

Ⅱ. 선행연구

2.1 지능형 지휘통제 시스템

지휘결심 시스템은 군사 의사결정 과정에 고려되는 다양한 요소들을 통합하여 임무 수행의 효율성을 높이기 위해 제안되었으며[11,12], 지휘 및 통제만을 구성 요소로 고려하는 C2 시스템을 시작으로 발전되었다[13].

이는 전장 내 신속한 의사소통이 중요해짐에 따라 통신 및 정보 기술과의 결합을 추가적으로 고려한 C3I 시스템이 제안되었다[14]. 이후 전장 환경이 정보 및 네트워크 중심전으로 변화함에 따라 데이터 처리 능력이 중요한 요소로 대두되었다. 이에, 기존의 C3I 시스템은 신속한 정보 처리를 위한 컴퓨터 요소가 추가된 C4I 시스템으로 확장되었다[15]. 최근에는 복잡한 환경에서 고성능의 분석 및 의사결정 지원을 위한 인공지능 기반의 지능형 의사결정 시스템이 제안되고 있다[1,16].

지능형 의사결정 시스템은 전장 상황에 대한 지능형 전장 분석 및 지휘관 의사결정 지원을 주요 요소로 고려한다. 전장 분석의 경우, 드론 및 위성을 통해 수집된 데이터를 기반으로 전장 환경 분석 자동화를 진행한다[17,18]. 이때, 해당 시스템의 핵심 요소로는 전장 내 표적의 종류 판별 정확도 최대화[16] 및 인지 과정의 지연시간 최소화가 고려된다[18]. 대표적인 방법으로는 경량화 모델 기반의 저지연 시스템 연구와[18] 고성능 인지 모듈 도입을 위한 Transformer 모델 기반의 연구가 활발히 진행되고 있다[17]. 지휘관 의사결정 지원 모듈에서는 인지 모듈에서의 분석 결과를 기반으로 실질적인 의사결정을 수행하며, 복잡한 환경에서 성공적 의사결정이 가능한 강화학습 기반 방식이 유망한 방법론으로 고려되고 있다[4,5].

2.2 군사 의사결정 시스템

군사 환경에서는 제한된 시간, 자원 등 복잡한 제약조건과 높은 불확실성이 결합된 영역에서 성공적으로 무기 자원을 할당해야 한다. 무기 - 표적 할당(Weapon - Target Assignment; WTA)은 주어진 표적들에 대해 효율적인 화력 조합을 선택하여 군사적 이익을 극대화하는 것을 목표로 하는 대표적 의사결정 문제로, 지휘통제와 자원관리 연구의 핵심 과제로 다뤄져 왔다.

초기 WTA 연구들은 규칙 기반 모델이나 수리적 최적화 기법, 휴리스틱 접근을 고려하였다[19-22]. [19] 및 [20]에서는 간소화된 전장 환경 내에서 최적화 기법인 분기 한정법과 정수계획법을 결합하여 문제를 해결하였으며, [21]과 [22]는 유전 알고리즘, 인공 벌 군집 알고리즘(Artificial Bee Colony ; ABC) 등의 방법을 적용하여 보다 복잡한 탐색공간을 갖는 WTA 문제의 탐색 효율을 향상시켰다. 최근에는 동적인 환경에서의 기존 방법론의 어려움[23,24]을 해결하기 위해 인공 신경망을 도입하고 있으며, 심층 강화학습이 대표적으로 고려되고 있다[4,25].

심층 강화학습 기반 WTA 연구는 기존 접근 방법들 대비 높은 적응성과 효율성을 보이고 있다[5,25,26]. [5]에서는 포인터 네트워크 기반 강화학습 구조를 통해 WTA 문제를 순차적 할당 문제로 재정의함으로써, 기존 휴리스틱 및 최적화 기반 방법 대비 의사결정 계산 시간을 크게 단축하면서도 유사한 할당 품질을 달성하였다. [4]은 표적 피해 최대화 및 자원 소모 최소화의 복합 목표 달성을 위해 강화학습 알고리즘 DDPG 기반의 DNPE(Deep Neighborhood Policy Exploration) 전략을 통해 기존 방법론 대비 개선된 의사결정 능력을 보였다.

이처럼 강화학습 기반 방법론은 계산 효율성 및 목표 기반 적응성을 동시에 확보한다. 하지만 군사 환경에서는 실시간 시행착오 기반의 기존 강화학습 방법론에서 잘못된 의사결정이 큰 위험을 초래할 수 있기 때문에 보다 안전한 학습 방법론이 요구된다. 이에 본 연구에서는 사전 수집 데이터에 대한 모방학습 항을 추가로 고려함으로써, 학습 과정에서의 안정성을 높인다.

2.3 데이터 기반 강화학습

기존 온라인 강화학습에서는 학습 데이터 수집 과정에서 환경과의 상호작용을 요구하는 비용이 높은 데이터 수집 방법을 고려하였다. 이는 수집 데이터의 재활용이 가능한 리플레이 버퍼의 도입으로 완화되었지만[27,28], 여전히 높은 수집 비용이 요구된다. 이에 사전 수집 데이터의 활용을 통해 학습의 효율성을 높이기 위한 강화학습 연구가 진행되고 있다[6,8,9].

[8]에서는 사전 수집된 데이터와 리플레이 버퍼 데이터를 단순히 학습에 함께 활용하는 RLPD 방법론을 제안하였다. 해당 방법론은 구현이 용이한 동시에, 다양한 의사결정 환경에서 학습 효율성을 크게 증가시켰다. [9]에서는 전문가에 의해 수집된 데이터에 대한 모방학습 항을 도입하는 CoL 방법을 도입하였다. 해당 방법론은 학습 모델이 데이터셋과 유사한 행동을 결정하도록 유도함으로써 빠르게 모델을 개선할 수 있다는 장점을 갖는다. 하지만 해당 방법론은 데이터에 대한 모방을 고려하므로, 고품질의 전문가 데이터에만 한정적으로 적용되어야 한다는 어려움이 있다.

본 연구에서는 사전 수집 데이터 중 우수한 행동 정보만을 선별적으로 모방하는 방법론의 도입을 통해[6] 다양한 수준의 개체로부터 수집된 데이터에도 효율적인 학습을 할 수 있는 의사결정 시스템을 구축한다.

Ⅲ. 시스템 설정

본 연구에서는 지휘관 요망효과를 성공적으로 달성할 수 있는 탄종 및 사용량 의사결정을 고려한다. 이에,

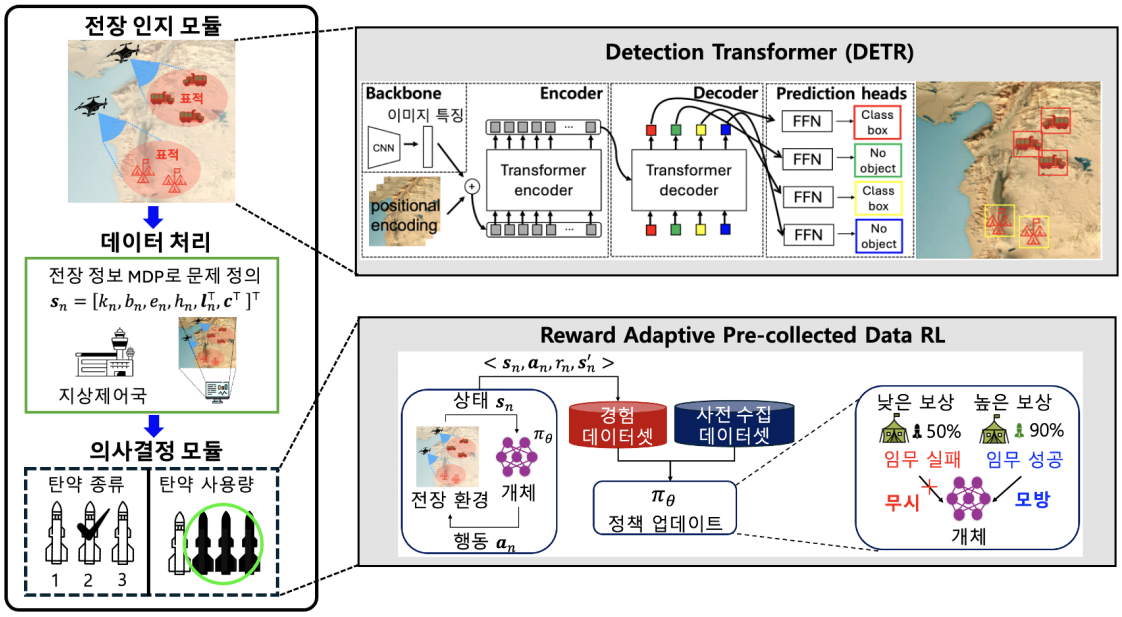


그림 1. 지휘결심 지원 시스템
Fig. 1. Military decision support system

본 장에서는 지휘결심 지원을 위한 화력운용 의사결정 문제를 해결하기 위한 시스템 구조 및 구성요소를 설명한다. 제안하는 지휘결심 시스템은 그림 1과 같이 전장 인지, 통신 및 데이터 처리, 그리고 의사결정 모듈로 구성된다.

3.1 전장 인지

전장 인지 모듈은 정찰 드론, 지상 레이더 등 다양한 센서를 통해 표적의 위치, 종류, 규모 등의 정보를 탐지한다. 입력 영상 x는 RT-DETR기반 객체 탐지 모델을 통해 처리되며[29], 특징 맵 F에서 Transformer 인코더-디코더 구조를 거쳐 각 객체의 클래스 y_i와 위치 m_i가 예측된다. 인지 모듈에서 탐지된 결과 (y_i, m_i)는 표적의 종류와 위치 정보를 포함하며, 이후 단계에서 상태 정보 s_n의 일부로 사용된다. 즉, 전장 인지 모듈은 복잡한 관측 데이터를 구조화하여 의사결정에 필요한 형태로 변환하는 역할을 수행한다.

3.2 통신 및 데이터 처리

인지 단계에서 생성된 객체 정보 (y_i, m_i)는 통신망을 통해 지상 제어국으로 전송된다. 지상 제어국은 표적 인지 결과로 표적 부대의 종류 k_n, 표적의 상태에 따른 방호도 b_n, 표적의 전투력 h_n과 아군 부대의 정보들을 통합하여 전장 환경의 상태 정보 s_n으로 매핑한다. 이 상태 정보는 강화학습 기반 의사결정 모듈의 입력으로 사용되며, 전장 상황의 복잡한 상호관계를 정량적으로 반영한다.

3.3 강화학습 기반 의사결정

의사결정 모듈은 상태 정보 s_n을 입력으로 받아, 각 표적에 대해 적절한 탄종과 사용량으로 정의되는 행동 a_n을 산출한다. 각 행동은 전장 제약조건을 만족해야 하며, 선택된 행동에 따라 다음 상태 s'_n와 보상 $r_n = R(s_n, a_n, s'_n)$이 결정된다. 보상 함수는 지휘관의 요망효과 달성 정도, 탄약 운용의 효율성, 자원 소모의 균형 등을 함께 고려하며, 정책 $\pi_\theta(a_n|s_n)$은 심층 신경망으로 근사된다. 학습 과정에서 정책은 기대 누적 보상 $\mathbb{E}[\sum_n \gamma^n r_n]$을 최대화하도록 갱신되며, 다양한 환경 변화에도 안정적인 의사결정을 수행할 수 있도록 학습된다. 본 연구에서는, 효율적인 강화학습을 위해 사전 수집된 데이터 중 높은 보상을 갖는 행동에 대한 선별적인 모방을 고려하는 RAPD-RL 알고리즘을 도입한다[6].

3.4 시스템 목표 및 제약조건

본 시스템의 목표는 지휘관이 지정한 요망효과를 성공적으로 달성하는 것이다. 본 연구에서는 지휘관의 요망효과를 표적에 가해야 할 목표 피해량 e_n으로 표현하며 지휘관이 바라는 표적의 목표 잔여 전투력은 h_{e_n}으로 정의한다. 이때, 실제 화력운용으로 인한 표적에 대한 총 피해량 p_n이 반영되었을 때 표적 전투력 h'_n이

h_{e_n}에 도달하도록 정책을 학습하는 것을 목표로 한다.

동시에 실제 운용에서는 다양한 제약조건이 존재하므로, 정책 학습은 해당 제약조건을 준수하면서 목표 피해량을 달성하도록 유도되어야 한다. 예를 들어, 특정 표적에 대해 사용이 제한된 탄종이나 일부 화기가 특정 탄약만 운용할 수 있는 제약, 아군의 잔여 자원 $\boldsymbol{l}_n$의 한계, 표적의 종류와 방호 상태 등이 고려될 수 있다.

Ⅳ. 보상 적응형 강화학습 기반 군사 지휘결심 지원 시스템

본 장에서는 제안하는 지휘결심 지원 시스템의 핵심 요소를 설명한다. 먼저, 전장 인지를 위해 적용한 RT-DETR 기반 객체 탐지 모델을 설명한다. 이어서, 인지 결과를 입력으로 하여 본 논문의 핵심인 화력운용 문제를 MDP로 정의하고, 상태·행동·보상 구조를 구체화한다. 마지막으로, RAPD-RL 알고리즘 기반 의사결정 모델을 설명한다.

4.1 DETR 기반 인지 및 데이터 처리

전장 인지 단계에서는 정밀한 객체 탐지 및 추론 속도 보완을 위해 RT-DETR 모델을 활용한다[29]. RT-DETR은 객체 탐지를 집합 예측(set prediction) 문제로 정식화하며, Transformer 기반 인코더 - 디코더 구조를 통해 입력 영상 내 객체의 위치와 범주를 동시에 예측한다.

입력 영상 $x \in \mathbb{R}^{H\times W\times 3}$은 CNN 백본을 통해 특징 맵 $F \in \mathbb{R}^{M\times H'\times W'}$으로 변환된다. 여기서 H, W는 원본 영상의 세로, 가로 크기, M는 출력 채널 수, H', W'는 downsampling된 공간 해상도를 의미한다. 이 특징 맵은 위치 인코딩이 추가된 뒤 Transformer 인코더에 입력되며, self-attention을 통해 전역적인 문맥 정보를 통합한다.

$$z = Encoder\left(F + E_{pos}\right), \quad z \in \mathbb{R}^{O\times d}$$

여기서 E_{pos}는 위치 인코딩, $O = H' \times W'$은 토큰 개수, d는 임베딩 차원을 의미한다. 이후 디코더는 고정된 개수의 object query q_i를 입력으로 받아 각 query가 하나의 잠재적 객체를 담당하도록 한다. 디코더 출력은 선형 헤드를 통해 각 표적 후보의 클래스 확률 $\hat{y}_i$와 bounding box의 위치 $\hat{m}_i$를 출력한다.

$$(\hat{y}_i, \hat{m}_i) = Decoder(q_i, z)$$

모델 학습은 Hungarian algorithm 기반의 일대일 매칭을 통해 수행되며, 손실 함수는 분류 손실과 회귀 손실의 조합으로 정의된다.

$$\mathcal{L}_{DETR} = \mathcal{L}_{cls} + \mathcal{L}_{box}$$

여기서 $\mathcal{L}_{cls}$는 cross-entropy 기반의 분류 손실, $\mathcal{L}_{box}$는 bounding box 회귀 손실을 의미한다. 탐지 결과인 ($\hat{y}_i$, $\hat{m}_i$)는 표적과 아군의 정보로 활용되며, 지상제어국에서 의사결정 모듈의 상태(state) 정보 $\boldsymbol{s}_n$로 처리된다.

4.2 Markov Decision Process 모델 정의

본 연구에서는 강화학습 기반 의사결정을 위해 화력운용 문제를 MDP로 모델링하였다. MDP는 튜플 <S, A, T, R, γ>로 정의되며, 여기서 S는 전장 환경의 상태 공간, A는 행동 공간, $T(\boldsymbol{s}'_n|\boldsymbol{s}_n, \boldsymbol{a}_n)$는 상태 전이 확률, $R(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}'_n)$은 보상 함수, $\gamma \in [0,1)$는 시간에 따른 감가율(discount factor)을 의미한다. 학습의 주체인 아군 부대는 개체로서 정의되며, 특정 전장 상태 $\boldsymbol{s}_n$에서 무기 선택 행동 $\boldsymbol{a}_n$를 수행하고 다음 상태 $\boldsymbol{s}'_n$에 도달하여 보상 $r_n = R(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}'_n)$을 획득하게 된다. 이때, 개체는 보상을 최대화하는 방식으로 의사결정 정책을 학습하게 된다.

4.2.1 상태 정보(state)

전장 환경의 상태 정보 $\boldsymbol{s}_n$는 n번째 에피소드에서 모든 정보를 의미하며, 다음과 같이 정의된다.

$$\boldsymbol{s}_n = [k_n, b_n, e_n, h_n, \boldsymbol{l}_n^{\top}, \boldsymbol{c}^{\top}]^{\top} \tag{1}$$

수식 (1)에서 $k_n \in \{k_{n,1}, \cdots, k_{n,K}\}$는 n번째 에피소드에 대한 K개의 표적 부대 종류, $b_n \in \{b_{n,1}, \cdots, b_{n,B}\}$는 B개의 표적 상태에 따른 방어도 상수, e_n은 n번째 에피소드에 대한 지휘관의 요망효과, h_n는 표적 부대의 전투력, $\boldsymbol{l}_n = [l_{n,1,1}, \cdots, l_{n,1,W}, \cdots, l_{n,U,1}, \cdots, l_{n,U,W}]^{\top}$는 n번째 에피소드의 아군 부대 종류와 무기 종류에 따른 잔여 탄수 벡터, $\boldsymbol{c} = [c_{1,1}, \cdots c_{1,W}, \cdots, c_{U,1}, \cdots c_{U,W}]^{\top}$는 아군 부대의 종류와 무기 종류에 따라 일정 기간 가용 보급을 고려해 할당될 수 있는 탄약 보급률 벡터를 의미한다.

4.2.2 행동 정보(action)

아군 부대의 n번째 에피소드 내 행동 $\boldsymbol{a}_n \in A$는 다음과 같이 정의된다.

$$\boldsymbol{a}_n = [a_n^{ammo}, a_n^{cost}]^{\top} \quad (2)$$

수식 (2)에서 $a_n^{ammo} \in \{1, 2, \dots, W\}$은 탄종 선택 행동으로 아군이 선택 가능한 W개의 탄약 종류 중 하나를 선택하는 행동을 한다. $a_n^{cost} \in [0, 1)$는 선택한 무기에 대한 탄약 사용량을 결정하는 행동이다. 이때, w무기에 대한 실제 사용한 탄약 수는 탄약 보급률인 $c_{u,w}$를 고려하여 다음과 같이 정의된다.

$$c_{u,w} \times a_n^{cost} \quad (3)$$

즉, 개체는 최대 $c_{u,w}$만큼의 탄약을 사용할 수 있다. 수식 (3)에 의해 실제 사용한 탄약을 기반으로 n번째 에피소드에서 사용 후 남은 탄약 수 $l'_{n,u,w}$는 보유 탄약 수 $l_{n,u,w}$에서 사용한 탄약 수의 차이로 다음과 같이 정의된다.

$$l'_{n,u,w} = \begin{cases} l_{n,u,w} - \left(c_{u,w} \times a_n^{cost}\right), & a_n^{ammo} = w \\ l_{n,u,w}, & otherwise \end{cases} \quad (4)$$

아군 행동에 의해 발생하는 피해량 $p_{n,u,w}$은 선택된 무기 종류와 사용 탄약 수에 따라 결정된다.

$$p_{n,u,w} = f\left(k_n, l_{n,u,w}, b_n, \boldsymbol{a}_n\right) \quad (5)$$

수식 (5)에서 $f(\cdot)$는 탄종 특성과 표적의 방호 상태를 반영한 피해 함수이며, 탄약 종류, 탄종별 살상 반경, 표적의 방어도 상수 등이 반영된다.

표적에 가한 총 피해량 p_n은 표적에 할당된 아군들의 피해량을 합산하여 계산된다.

$$p_n = \sum_{u=1}^{U} p_{n,u,w} \quad (6)$$

최종적으로 표적의 전투력은 피해량만큼 감소하며, 에피소드 종료 시 표적 부대의 전투력은 다음과 같이 정의된다.

$$h'_n = h_n - p_n \quad (7)$$

수식 (7)에서 h_n은 에피소드 시작 시 표적의 전투력이며, h'_n는 피해 반영 이후의 전투력을 나타낸다.

4.2.3 보상 함수(reward)

아군 부대의 n번째 에피소드에서 보상 r_n은 r_n = R $(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}'_n)$로 정의하며, 다음과 같이 보상항 및 처벌항의 선형 결합으로 이뤄진다.

$$R(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}'_n) = \eta_1 R_{n,1} + \eta_2 R_{n,2} \quad (8)$$

수식 (8)에서 η_1, η_2는 각 항에 대한 계수를 나타내고, $R_{n,1}, R_{n,2}$는 임무 달성을 위한 항과 지휘관 선호도 반영을 위한 항을 의미한다.

임무 달성 보상항 $R_{n,1}$은 지휘관이 설정한 요망효과 e_n와 실제 피해량 p_n의 차이를 기반으로 정의된다.

$$R_{n,1} = \begin{cases} \max\left(\dfrac{p_n - e_n}{\omega\zeta} + 1, R_1^{\min}\right), & p_n < e_n \\ \max\left(-\dfrac{p_n - e_n}{2\omega\zeta} + 1, R_1^{\min}\right), & p_n \geq e_n \end{cases} \quad (9)$$

여기서 ζ는 피해도 마진을 나타내며, 화력 결정으로 인한 피해량이 요망효과에 정확히 일치할 경우($p_n = e_n$) 가장 높은 보상이 부여되며, 초과 달성 또는 미달성의 경우에는 보상이 점진적으로 감소한다. 특히 미달성에 대한 보상 감소율은 초과 달성의 경우보다 두 배로 크게 설정하여, 요망효과를 달성하지 못하는 상황에 더 큰 페널티를 부여하였다. 또한 학습의 안정성을 위해 마진 민감도 $\omega \in (0,1]$를 도입하여 개체가 고려하는 마진의 범위를 보수적으로 설정하도록 정의하였으며, 최소 보상 $R_1^{\min}$을 통해 페널티의 크기가 과도하게 증가하지 않도록 제한하였다.

본 연구에서는 지휘관의 전술적 의도와 운용 제약을 정책 학습 과정에 반영하기 위해 화력운용 제약조건을 설정한다. 각 제약은 상태, 행동 쌍 $(\boldsymbol{s}_n, \boldsymbol{a}_n)$의 집합으로 정의되며, 제약을 위반하는 상태 - 행동 조합을 $\mathcal{C} \subseteq S \times A$로 표현한다. 즉, $(\boldsymbol{s}_n, \boldsymbol{a}_n) \in \mathcal{C}$이면 해당 상태에서의 행동이 제약을 위반한 경우로 간주된다. 이에 따라 탄종 선택 처벌항 $R_{n,2}$는 다음과 같이 정의된다.

$$R_{n,2} = \begin{cases} -1, & (\boldsymbol{s}_n, \boldsymbol{a}_n) \in \mathcal{C} \\ 0, & otherwise \end{cases} \quad (10)$$

즉, 제약을 위반한 경우 -1의 보상이 부여되며, 그렇지 않은 경우 0으로 설정된다.

결과적으로 제안된 보상 함수는 임무 달성도와 전술적 제약을 동시에 반영하여, 학습된 정책이 요망효과를 달성하면서도 운용 제약을 만족하도록 유도한다. 이러한 설계는 실제 운용 환경에서의 지휘결심 지원 시스템이 안정적이고 신뢰성 있는 정책을 학습하는 데 기여한다.

4.3 사전 수집 데이터를 활용한 보상 적응형 강화학습

효율적인 학습을 위해 본 연구에서는 온라인 상호작용으로 수집된 데이터뿐만 아니라 사전에 확보된 데이터셋을 병행하여 활용한다. 본 연구에서는 사전 데이터 활용하는 보상 적응형 강화학습 방법을 적용하였다.

제안하는 방식은 두 개의 버퍼를 병행하여 사용한다. 하나는 환경 상호작용을 통해 수집되는 온라인 버퍼 D_{on}이며, 다른 하나는 사전에 확보된 데이터셋을 저장하는 버퍼 D_{pre}이다. 학습 과정에서는 두 버퍼에서 균등하게 데이터를 샘플링하여 정책 갱신에 활용한다. 이를 통해 온라인 강화학습의 데이터 효율성을 유지하면서도, 사전 데이터셋이 제공하는 다양한 전장 상황을 반영할 수 있다.

모방 손실항(BC loss)에 가중치로 부여하여, 일정 기준 이상의 보상을 갖는 전이 데이터만을 모방 대상으로 삼는다. 즉, 고품질 데이터는 적극적으로 모방하고, 저품질 데이터는 배제함으로써 학습 안정성과 정책 성능을 동시에 확보한다. 정책의 학습 목적 함수는 다음과 같이 정의된다.

$$\pi^* = arg\max_{\pi} \mathbb{E}_{(\boldsymbol{s_n},\boldsymbol{a}_n,r_n)\sim D_{on}\cup D_{pre}} \Big[Q_\theta\big(\boldsymbol{s}_n,\pi(\boldsymbol{s}_n)\big) - \sigma(r_n-\epsilon)(\pi(\boldsymbol{s}_n)-\boldsymbol{a}_n)^2\Big]^2 \tag{11}$$

여기서 Q_θ 는 상태 - 행동 가치 함수, π는 정책, ε은 모방 기준 보상, $\sigma(\cdot)$는 ReLU 연산자로, 샘플의 보상이 기준치를 초과할 경우에만 모방 손실이 활성화된다. 이를 통해 학습 정책은 온라인 및 사전 데이터셋 모두에서 유효한 행동을 선택적으로 모방하게 된다. 이러한 보상 적응형 학습 전략은 학습 초기에 사전 데이터셋을 적극적으로 활용하여 빠른 수렴을 유도하고, 동시에 학습이 진행됨에 따라 온라인 데이터로부터 개선된 정책을 습득할 수 있도록 한다. 결과적으로 제안된 방식은 저품질 데이터셋이 포함되어 있더라도 강인한 학습 성능을 유지하며, 전장 환경과 같은 복잡한 문제에서 안정적이고 효율적인 정책 학습을 가능하게 한다.

Ⅴ. RT-DETR 기반 인지 실험 설정 및 분석

본 절에서는 제안된 지휘결심 지원 시스템의 인지 모듈 성능 검증을 위한 성능 지표와 분석 결과를 제시한다.

5.1 성능평가 지표 및 평가 데이터

본 연구에서는 제안 시스템의 성능 평가를 위해 탐지 성능 및 지연 시간을 고려한다. 또한, 다중 클래스 탐지 능력과 실제 운용 성능을 모두 평가한다.

표 1. 제안 방법 인지 성능
Table 1. Perception performance of proposed method

	In-domain	Cross-domain
mAP (%)	99.99	86.68
Latency (*ms*)	129*ms*	

5.1.1 객체 탐지 시스템 평가 지표

- 평균 객체 인식률(mAP): 탐지된 객체의 클래스 개수 Y에 대한 평균 정밀도(AP)를 기반으로 산출하며, 인지 모듈이 다양한 표적을 얼마나 정확히 식별하는지를 나타낸다.

$$mAP = \frac{1}{Y}\sum_{i=1}^{Y} AP_i \times 100\ (\%)$$

- 평균 처리속도(Latency): X개의 입력 영상을 고려했을 때, 영상 i에 대해 객체 인식이 시작된 시각 B_i와 완료된 시각 A_i의 차이를 기반으로 평균 처리속도를 계산한다.

$$\text{Latency} = \frac{1}{X}\sum_{i=1}^{X}(A_i - B_i)$$

5.1.2 성능 평가 환경

- In-domain: 학습 중 경험한 데이터에 대한 인지 성능을 제공하며, 모델의 학습 데이터 및 평가 데이터가 서로 같은 데이터셋에서 분리된 환경을 의미한다. 해당 데이터셋은 자체 수집된 군 장비의 실제 이미지를 포함하여 시뮬레이션 이미지 및 모형 장비 이미지로 구성된다.
- Cross-domain: 실제 운용 환경에서의 탐지 성능을 제공하며, 학습 데이터와 평가 데이터가 수집된 환경이 상이하다는 특성이 있다. 학습 데이터로는 MS COCO[32] 및 Objects 365[33]를 고려하며, 소수의 자체 수집 데이터를 포함한다. 성능 평가를 위한 데이터는 실제 운용 장비가 수집한 이미지만을 고려하였다.

5.2 성능 분석 및 평가

표 1은 본 연구에서 고려하는 성능 평가 모델에 대한 mAP 및 평균 처리속도 결과로 제안 시스템의 인지 성능을 나타낸다. 해당 결과를 통해 In-domain 환경을 고

려할 경우 평균 정확도(mAP)가 100%에 가까운 성능을 달성하며, Cross-domain 환경에서도 86.68%의 높은 성능을 보이는 것을 확인할 수 있다. 이는 다양한 전장 표적 상황에서 제안된 인지 모듈이 안정적으로 작동하며, 핵심 표적 분류에 있어서는 고성능의 식별 능력을 제공함을 의미한다. 또한, 평균 처리속도는 129ms로 측정되어, 초 단위의 빠른 반응이 요구되는 전술 상황에서 실시간 의사결정을 지원할 수 있는 수준임을 확인하였다.

종합적으로 제안된 시스템은 임무 목표 달성의 효과와 표적 인식 정확도 및 실시간성 보장, 제약조건 준수 측면에서 모두 우수한 성능을 달성하는 것을 확인할 수 있다.

Ⅵ. 강화학습 기반 판단 실험 설정 및 분석

본 연구에서는 강화학습 기반 의사결정 모델 학습을 위해 단일 시점 의사결정 에피소드 N= 3000 회에 대한 학습을 진행하였다. 아군 부대의 수는 U = 4로 설정하였으며, 각 부대는 최대 W = 4 종류의 탄종을 운용할 수 있다. 표적의 종류는 K = 4로 구성되며, 각 표적 유형별로 방호 상태에 따라 두 가지 수준의 방어도 B = 2가 정의된다. 본 연구에서는 피해 산정을 위해 기존 군사 교전 모델에서 널리 사용되는 두 가지 대표적 피해 함수, Carleton 모델[30]과 Cookie-Cutter 모델[31]을 기반으로 피해 함수를 설계하였다.

6.1 의사결정 정책 비교 알고리즘 및 평가 지표

본 절에서는 의사결정 모델 평가를 위한 비교 알고리즘 및 평가 지표를 제공한다.

6.1.1 비교 알고리즘

제안한 MDP 기반 강화학습 정책과의 성능 비교를 위해, 강화학습 알고리즘(RL-based)과 다중 목표 최적화 휴리스틱 알고리즘을 고려하였다.

1) RL-based: 심층 강화학습 기반의 정책 비교를 위해 액터 - 크리틱 알고리즘과 사전 수집 데이터를 활용하는 방법을 고려한다.

- DDPG[27]: 결정론적 정책(deterministic policy)을 사용하는 알고리즘으로, 정책 기반 DPG에 DQN 구조의 크리틱 네트워크를 결합하여 연속적 행동 공간에서 학습이 가능하다.
- TD3[28]: 두 개의 크리틱 네트워크를 사용하여 작은 Q값을 선택하는 double Q 기법을 적용하고, 액터 업데이트를 지연시키는 전략을 통해 DDPG의 과대 추정 문제를 완화하고 안정성을 높였다.
- RLPD[8]: 온라인 버퍼와 사전 수집 버퍼를 동시에 사용하는 방법으로, BC 정규화 항을 포함하지 않고 학습한다. 이를 통해 BC 손실항의 효과를 분석하기 위한 기준선으로 활용된다.
- CoL[9]: 온라인 버퍼와 사전 수집 버퍼를 모두 사용하는 방법으로, BC 정규화 항을 포함한다. 단, 선택적 모방 조건을 두지 않아 데이터 품질이 낮을 경우 성능 저하가 발생할 수 있다. CoL은 다음과 같은 두 가지 방식으로 구분된다.
  - CoL-pre: 사전 데이터셋만 모방 대상으로 사용하는 경우
  - CoL-all: 온라인 데이터와 사전 데이터셋을 모두 모방 대상으로 사용하는 경우

2) Heuristic: 지휘관의 요망효과 달성과 자원 효율성을 동시에 고려하기 위해, 다중 목표 최적화를 수행할 수 있는 다음과 같은 휴리스틱 알고리즘을 비교 대상으로 포함하였다.

- MOGA (Multi-Objective Genetic Algorithm)[21]: 개체군 기반 탐색을 통해 다양한 해를 병렬적으로 탐색하며, 복잡한 다중 목표 최적화 문제에서 전역 최적해에 수렴할 가능성을 높인다.
- MOABC (Multi-Objective Artificial Bee Colony)[22]: 탐색과 탐색 강화를 균형 있게 수행하는 알고리즘으로, 초기 해 공간을 광범위하게 탐색한 뒤 탐색 강화를 통해 해의 품질을 개선한다.

6.1.2 데이터셋 및 평가 지표

본 연구에서 사용된 데이터셋은 군사 시뮬레이션 환경에서 규칙 기반의 에이전트로 수집한 전이 데이터로 구성되며, 다음과 같이 구분된다.

- Final: 전문가 수준의 정책을 반영하는 데이터셋으로써, 실제 전장에서의 합리적인 지휘관 의사결정에 가까운 행동 데이터를 포함한다.
- Final+Medium: 부분적인 전문가 수준의 정책을 반영하는 데이터셋이다. 이는 학습 과정에서 나타나는 불완전한 의사결정을 포함하여, 다양한 수준의 의사결정 전략을 반영한다.

본 연구의 사전 수집 데이터셋은 지휘관의 의사결정 데이터를 반영하기 위해, 품질이 낮은 무작위 정책 데이터는 제외하고 Final 및 Medium 수준의 고품질 데이터만을 채택하였다.

성능 비교를 위한 테스트 시나리오는 모든 표적 조합

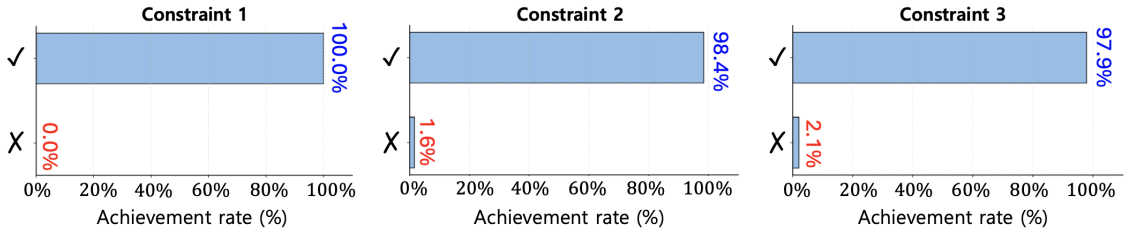


그림 2. 제안 방법의 전장 제약조건 달성 여부
Fig. 2. Achievement of the constraints of the proposed method

표 2. 제안 방법 의사결정 성능
Table 2. Decision-making performance of proposed method

	Dataset	Proposed
Achievement	Final	90.5±3.3
	Final+Medium	81.6±4.5
Reward	Final	1.8±0.6
	Final+Medium	-0.4±4.3

과 요망효과 조합이 포함된 총 N_{test}개의 에피소드에 대해 평가를 진행하였다. 성능 평가 지표는 다음과 같다.

- 달성률(Achievement): 표적의 최종 전투력이 요망 전투력의 허용 마진 범위 내에 있을 때 해당 에피소드를 달성된 것으로 정의하며, 요망 전투력에 미치지 못하거나 초과 피해를 입힌 경우 달성되지 않은 에피소드로 간주한다.

$$g(n)=\begin{cases} 1, & 0 \le h_{e_n} - h'_n \le \varepsilon \\ 0, & \text{otherwise} \end{cases}$$

여기서, $g(n)$는 n번째 에피소드에 대한 달성 함수이며, 1은 요망 전투력 달성 에피소드, 0은 미달성 에피소드를 의미한다. 이때 달성률은 전체 N_{test}개의 에피소드 중 달성된 에피소드의 비율로 다음과 같이 계산한다. 해당 값이 클수록 성능이 우수함을 의미한다.

$$\text{Achievement} = \frac{1}{N_{test}} \sum_{n=1}^{N_{test}} g(n) \times 100 \ (\%)$$

- 평균 보상(Reward): 각 에피소드에 대해 누적되는 보상을 계산한다. 평균 보상은 총 N_{test}개의 에피소드에 대한 보상의 평균으로 정의된다.

$$\text{Reward} = \frac{1}{N_{test}} \sum_{n=1}^{N_{test}} r_n$$

6.2 성능 분석 및 평가

본 절에서는 제안한 방법의 임무 달성과 전장 제약조건 충족 여부를 중심으로 성능을 분석한다.

1) 의사결정 정책 성능: 표 2는 제안한 방법의 의사결정 정책 성능을 나타낸다. Final 데이터셋에서 평균 90.5%의 달성률을 기록하였으며, 이는 다양한 전장 조건에서 임무 목표를 안정적으로 달성할 수 있음을 의미한다. 또한, Final+Medium 데이터셋에서도 81.6%의 달성률을 유지하여, 중간 수준의 정책이 혼재된 데이터 상황에서도 강인한 성능을 발휘함을 확인하였다.

2) 전장 제약조건 달성 여부: 제안한 방법의 제약조건 충족 성능을 분석하기 위해 다음 세 가지 전장 제약조건을 고려한다.

- 제약조건 1: 특정 표적에 대해 사용이 금지된 탄종이 존재한다.
- 제약조건 2: 특정 표적에 대해 우선적으로 사용되어야 하는 탄종이 존재한다.
- 제약조건 3: 각 아군 부대에 대해 사용할 수 있는 탄종이 제한되어 있다.

그림 2는 제안한 방법으로 학습한 의사결정이 제약조건을 얼마나 충족하는지 분석한 결과이다. 제약조건 1의 경우, 특정 표적에 대해 특정 탄약의 사용이 금지되어야 하는데, 해당 조건이 100% 충족되어 불필요한 자원 낭비를 방지한 것을 확인할 수 있다. 제약조건 2에서는 특정 표적에 대해 특정 탄약 계열이 우선적으로 선택되어야 하는 상황을 평가하였으며, 제안된 정책은 98.4%의 높은 충족률을 보였다. 제약조건 3은 특정 아군 부대에 대해 제한된 탄약 종류만 사용할 수 있다는 제약이며, 97.9%로 해당 제약조건을 만족하였다.

종합적으로 제안된 시스템은 임무 목표 달성의 효과와 제약조건 준수 측면에서 모두 우수한 성능을 달성하는 것을 확인할 수 있다.

6.3 비교 알고리즘과의 의사결정 성능 비교

본 절에서는 제안된 보상 적응형 강화학습 기반 정책을 기존의 RL 기반 및 휴리스틱 기반 알고리즘들과 직

표 3. 알고리즘별 성능 비교
Table 3. Compare performance by algorithm

	Dataset	RL-based						Heuristic-based	
		Proposed	RLPD	CoL-pre	CoL-all	DDPG	TD3	MOGA	MOABC
Achievement	Final	90.5±3.3	86.7±5.3	77.7 ±11.7	77.8 ±11.3	74.7±4.6	79.2±3.8	56.7±0.1	66.2±0.1
	Final+ Medium	81.6±4.5	60.5 ±25.8	63.8 ±23.5	51.7 ±30.8				
Reward	Final	1.8±0.6	0.5±1.1	-12±2.7	-1.0±2.7	-1.3±2.5	-0.7±1.9	-5.8±0.1	-3.6±0.1
	Final+ Medium	-0.4 ±4.3	-4.7 ±5.7	-4.1 ±5.7	-7.9 ±9.0				

접 비교한다. 표 3은 Final 및 Final+Medium 데이터셋에서 각 알고리즘의 임무 달성률을 요약한 결과이다.

전통적인 RL 기반 접근법인 DDPG와 TD3는 초기 정책이 미숙할 경우 저품질 데이터가 대량으로 수집되어 학습 효율이 저하되는 문제가 존재하기 때문에 실제로 DDPG는 Final 데이터셋에서 약 74.7%, TD3는 79.2%의 달성률을 기록하여, 제안된 방법 대비 10% 이상 낮은 성능을 보였다. RLPD와 CoL 계열 알고리즘은 사전 수집 데이터를 활용하여 성능을 보완할 수 있으나, 데이터 품질에 민감하게 반응하는 한계를 드러냈다. 특히 RLPD는 86.7%의 성능을 기록하여 비교적 높은 달성률을 달성했으나, Final+Medium 데이터셋에서는 66.7%로 급격히 하락하여 데이터 혼합 상황에서의 강건성이 부족함을 확인할 수 있었다. CoL-pre와 CoL-all은 Final 데이터셋에서 각각 77.7%, 77.8% 수준의 성능에 그쳤고, Final+Medium 데이터셋에서도 성능이 급격하게 하락하는 것을 확인할 수 있다. 이에 비해 제안된 방법은 Final 데이터셋에서 90.5%, Final+Medium 데이터셋에서 81.6%를 달성하며, 고품질 데이터뿐만 아니라 중간 수준의 정책이 혼재된 데이터 상황에서도 일관된 성능을 유지하였다.

다중 목표 최적화 접근법인 MOGA와 MOABC는 전역 탐색을 통해 일정 수준의 해를 도출할 수 있었으나, 각각 56.7%, 66.2%의 달성률에 그쳐 강화학습 기반 접근법 대비 성능이 현저히 낮았다. 이는 휴리스틱 방법이 사전에 정의된 규칙과 탐색 전략에 크게 의존하기 때문에, 동적으로 변화하는 전장 조건이나 지휘관의 성향 반영에 유연하게 대응하기 어렵다는 점을 의미한다.

종합적으로, 제안된 방법은 RL 기반 알고리즘 대비 데이터 품질의 민감성을 완화하고, 휴리스틱 기반 접근법 대비 전장 환경 변화에 적응할 수 있는 유연성을 확보하였다. 이는 제안된 보상 적응형 강화학습 기법이 단순히 특정 조건에서만 효과적인 것이 아니라, 데이터 환경과 제약조건이 변화하는 다양한 상황에서도 강인한 성능을 유지할 수 있음을 실험적으로 확인할 수 있다.

Ⅶ. 결 론

본 논문에서는 보상 적응형 강화학습 기반 지휘관의 지휘결심 지원을 위한 시스템을 제안하였다. 제안된 시스템은 전장 인지, 통신, 판단으로 이어지는 end-to-end 절차를 기반으로 설계 되었으며, 특히 타격 방법 결정 단계에 강화학습을 적용하여 지휘관의 요망효과와 환경의 제약을 동시에 반영할 수 있도록 하였다. 이를 위해 MDP 모델을 정의하고, 사전 수집 데이터와 온라인 상호작용 데이터를 병행 활용하는 보상 적응형 모방 기법을 도입하였다. 실험 결과, 제안된 방법은 기존 RL 기반 및 휴리스틱 기반 접근법 대비 임무 달성률을 유의미하게 향상시켰으며, 환경의 제약조건 역시 높은 수준에서 충족하였다. 또한 인지 모듈과의 연계 실험을 통해 제안 시스템이 실제 전장 환경에서 요구되는 정확도와 처리속도를 동시에 만족할 수 있음을 확인하였다.

References

[1] C. Lee, et al., "Deep AI military staff: Cooperative battlefield situation awareness for commander's decision making," *J. Supercomputing*, vol. 79, no. 6, pp. 6040-6069, 2023.

[2] J. Han, et al., "Conceptual design of infrastructure and framework for a futuristic surveillance imagery fusion system," *J. KICS*, vol. 46, no. 9, pp. 1426-1439, 2021.

[3] S. Jin, et al., "A study on multiple reasoning technology for intelligent battlefield situational awareness," *J. KICS*, vol. 45, no. 6, pp. 1046-

1055, 2020.

[4] T. Li, et al., “An intelligent algorithm for solving weapon-target assignment problem: DDPG-DNPE algorithm,” *Computers, Materials & Continua*, vol. 76, no. 3, pp. 3499-3522, 2023.

[5] H. Na, et al., “Weapon-target assignment by reinforcement learning with pointer network,” *J. Aerospace Inf. Syst.*, vol. 20, no. 1, pp. 53-59, 2023.

[6] C. Eom, et al., “Selective imitation for efficient online reinforcement learning with pre-collected data,” *ICT Express*, vol. 10, no. 6, pp. 1308-1314, 2024.

[7] J. Lee, et al., “Deep reinforcement learning based weapon-target assignment to support military decision-making,” *J. KICS*, vol. 50, no. 6, pp. 884-895, 2025.

[8] P. Ball, et al., “Efficient online reinforcement learning with offline data,” *Int. Conf. Mach. Learn.*, 2023.

[9] V. Goecks, et al., “Integrating behavior cloning and reinforcement learning for improved performance in dense and sparse reward environments,” *Int. Conf. Autonomous Agents and MultiAgent Syst.*, 2020.

[10] Z. Huang, et al., “Efficient deep reinforcement learning with imitative expert priors for autonomous driving,” *IEEE Trans. Neural Networks and Learn. Syst.*, vol. 34, no. 10, pp. 7391-7403, 2022.

[11] C. Eom, et al., “A survey on weapon-target assignment for realistic battlefield environments: From exact algorithm to deep reinforcement learning,” *J. KICS*, vol. 50, no. 2, pp. 205-216, 2025.

[12] C. Han, et al., “A methodology of decision condition-based data modeling for constructing AI staff,” *J. ICS*, vol. 21, no. 1, pp. 237-246, 2020.

[13] B. Claverie and G. Desclaux, “C2-command and control: A system of systems to control complexity,” *Am. J. Manag.*, vol. 22, no. 2, pp. 45-63, 2022.

[14] J. Michaelis, et al., “Applying mission information requirements to value of information middleware,” *IEEE Military Commun. Conf.*, 2023.

[15] J. Kim, et al., “A study of artificial intelligence learning model to support military decision making: Focused on the wargame model,” *J. Korea Soc. Simulation,* vol. 30, no. 3, pp. 1-9, 2021.

[16] K. Park and S. Shim, “Intelligent counterforce allocation method using multi-agent reinforcement learning for ground operations,” *IEEE Access*, vol. 13, pp. 127009-127022, 2025.

[17] L. Chen, et al., “LCPD-DETR: A lightweight object detection model based on RT-DETR for military camouflaged personnel,” *J. Real-Time Image Process.*, vol. 22, no. 6, pp. 1-16, 2025.

[18] S. Ali, et al., “Computer vision-based military tank recognition using object detection technique: An application of the YOLO framework,” *IEEE Int. Conf. Advanced Innovations in Smart Cities*, 2023.

[19] D. Ahner and C. Parson, “Optimal multi-stage allocation of weapons to targets using adaptive dynamic programming,” *Optimization Lett.*, vol. 9, pp. 1689-1701, 2015.

[20] Y. Lu and D. Chen, “A new exact algorithm for the weapon-target assignment problem,” *Omega*, vol. 98, p. 102138, 2021.

[21] C. Wang, et al., “Multi-objective optimization of weapon target assignment based on genetic algorithm,” *Int. Conf. Computer, Internet of Things and Control Eng.*, 2021.

[22] H. Xing and Q. Xing, “An air defense weapon target assignment method based on multi-objective artificial bee colony algorithm,” *Computers, Materials & Continua*, vol. 76, no. 3, pp. 2685-2705, 2023.

[23] Y. Zhao, et al., “Multi-weapon multi-target assignment based on hybrid genetic algorithm in uncertain environment,” *Int. J. Advanced Robotic Syst.*, vol. 17, no. 2, 2020.

[24] X. Change, et al., “Adaptive large neighborhood search algorithm for multi-stage weapon target assignment problem,” *Computers & Industrial Eng.*, vol. 181, p. 109303, 2023.

[25] J. Liu, et al., "Intelligent air defense task assignment based on hierarchical reinforcement learning," *Frontiers in Neurorobotics*, vol. 16, p. 1072887, 2022.
[26] T. Lillicrap, et al., "Continuous control with deep reinforcement learning," *Int. Conf. Learn. Representations*, 2016.
[27] S. Fujimoto, et al., "Addressing function approximation error in actor-critic methods," *Int. Conf. Machine Learn.*, 2018.
[28] Y. Zhao, et al., "DETRs beat YOLOs on real-time object detection," *IEEE/CVF Conf. CVPR*, 2024.
[29] S. Moon, "Weapon effectiveness and the shapes of damage functions," *Defence Technol.*, vol. 17, no. 2, pp. 617-632, 2021.
[30] S. Dillenburger, et al., "Pareto-optimality for lethality and collateral risk in the airstrike multi-objective problem," *J. Operational Res. Soc.*, vol. 70, no. 7, pp. 1051-1064, 2019.
[31] T. Lin, et al., "Microsoft COCO: Common objects in context," *Eur. Conf. Computer Vision*, 2014.
[32] S. Shao, et al., "Objects365: A large-scale, high-quality dataset for object detection," *IEEE/CVF Int. Conf. Computer Vision*, 2019.

이 재 휘 (Jaehwi Lee)

2024년 2월 : 숭실대학교 전자정보공학부 IT융합전공 학사
2024년 3월~현재 : 숭실대학교 지능형반도체학과 석사과정
<관심분야> 인공지능, 강화학습, 지휘통제체계, 지능형지휘결심
[ORCID:0009-0001-8014-6493]

엄 찬 인 (Chanin Eom)

2022년 8월 : 숭실대학교 전자정보공학부 IT융합전공 학사
2022년 9월 : 숭실대학교 지능형반도체학과 석사과정
2025년 9월~현재 : 숭실대학교 지능형반도체학과 박사과정
<관심분야> 강화학습, 인공지능, 지능형지휘결심, 자율주행
[ORCID:0009-0005-6340-6635]

김 찬 (Chan Kim)

2023년 5월~현재 : 코난테크놀로지 인공지능연구소 전임연구원
<관심분야> Multimodal AI, Computer Vision, Large Language Models, MLOps
[ORCID:0009-0001-5412-0214]

김 경 수 (Kyeongsoo Kim)

2023년 8월~현재 : 코난테크놀로지 인공지능연구소 전임연구원
<관심분야> 강화학습, 인공지능, 지능형지휘결심, 비전인식
[ORCID:0009-0005-6453-0240]

이 형 도 (Hyeongdo Lee)

2021년 8월~현재 : 코난테크놀로지 인공지능연구소 책임연구원
<관심분야> 강화학습, 인공지능, 자율에이전트, 자율주행
[ORCID:0009-0009-5495-0873]

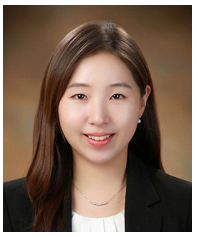
강 현 수 (Hyunsu Kang)

2016년 9월~현재 : 코난테크놀로지 인공지능연구소 이사

<관심분야> 강화학습, 지능형 지휘결심, 인공지능, 객체인식

[ORCID:0009-0001-5184-0259]

권 민 혜 (Minhae Kwon)



2011년 8월 : 이화여자대학교 전자정보통신공학과 학사

2013년 8월 : 이화여자대학교 전자공학과 석사

2017년 8월 : 이화여자대학교 전자전기공학과 박사

2017년 9월~2018년 8월 : 이화여자대학교 전자전기공학과 박사 후 연구원

2018년 9월~2020년 2월 : 미국 Rice University, Electrical and Computer Engineering, Postdoctoral Researcher

2020년 3월~2025년 2월 : 숭실대학교 전자정보공학부 IT융합 전공 조교수

2025년 3월~현재 : 숭실대학교 전자정보공학부 IT융합 전공 부교수

<관심분야> 강화학습, 지능형지휘결심, 자율주행, 모바일네트워크, 연합학습, 계산신경과학

[ORCID:0000-0002-8807-3719]

Appendix

표 A. 표기법 정의
Table A. Notation declaration

표기법	정의	표기법	정의
S	상태 공간	$\boldsymbol{s}_n$	상태
A	행동 공간	$\boldsymbol{a}_n$	행동
$T(\boldsymbol{s}'_n \mid \boldsymbol{s}_n, \boldsymbol{a}_n)$	상태 전이 확률	$R(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}'_n)$	보상 함수
γ	감가율	n	특정 에피소드
U	아군 부대의 수	a_n^{ammo}	아군 무기 선택 행동
W	아군 무기의 수	k_n	n번째 에피소드의 표적 부대 종류
b_n	n번째 에피소드의 표적 상태에 따른 방어도 상수	$\boldsymbol{c}$	아군 부대의 종류와 무기 종류에 따른 탄약 보급률 벡터
a_n^{cost}	선택한 부대와 무기에 대한 탄약 사용 비율 결정 행동	$\boldsymbol{l}_n$	n번째 에피소드에서 아군 부대와 무기 종류에 따른 잔여 탄수 벡터
K	표적 부대 종류의 수	h_{e_n}	n번째 에피소드에서 요망하는 표적 부대의 최종 전투력
B	표적 상태의 수	$g(n)$	n번째 에피소드에 대한 달성 함수
e_n	n번째 에피소드의 요망효과	ζ	요망 전투력 마진
h_n	n번째 에피소드에서 표적 부대의 전투력	N	전체 에피소드 수
η	보상 함수의 각 항에 대한 계수	W	요망 전투력 마진 민감도
$p_{n,u,w}$	n번째 에피소드에서 아군 부대 u가 탄종 w를 선택하여 가한 피해량		