Textile & Leather Review

ISSN 2623-6281 | www.tlr-journal.com | 10.31881/TLR

A Deep Reinforcement Learning Based on Spatio-Temporal Model for Solving Weapon -Target Assignment

Zimo Zhu, Chuanqiang Yu, Junti Wang

How to cite: Zhu Z, Yu C, Wang J. A Deep Reinforcement Learning Based on Spatio-Temporal Model for Solving Weapon-Target Assignment. Textile & Leather Review. 2026; 9:4012-4033. https://doi.org/10.31881/TLR.2026.4012

How to link https://doi.org/10.31881/TLR.2026.4012

Published:25 April 2026

A Deep Reinforcement Learning Based on Spatio-Temporal Model for Solving Weapon-Target Assignment

Zimo Zhu, Chuanqiang Yu*, Junti Wang

Department of Vehicle Engineering, Rocket Force University of Engineering, Xi'an 710025, Shaanxi, China
*yucq789@163.com

Article
https://doi.org/10.31881/TLR.2026.4012
Published 25 April 2026

ABSTRACT

The weapon–target assignment (WTA) problem plays a pivotal role in modern combat command and control systems because it directly influences operational effectiveness. Conventional solution methods, such as exact algorithms and heuristic approaches, often perform poorly in dynamic battlefield environments due to their limited ability to capture evolving engagement states. Deep reinforcement learning (DRL) offers a promising paradigm for sequential decision-making under such conditions; however, existing DRL-based WTA methods frequently overlook the spatio-temporal dependencies among combat entities, thereby constraining their adaptability and decision accuracy. To address this issue, this paper proposes a deep reinforcement learning with spatio-temporal modeling (DRLSTM) framework for dynamic WTA. Specifically, the framework integrates a graph convolutional network (GCN) to encode inter-entity spatial dependencies and a gated recurrent unit (GRU) to model temporal state evolution. This spatio-temporal architecture enables the agent to learn context-aware assignment policies from dynamic battlefield information. The policy is optimized using an actor–critic learning scheme. Experimental results demonstrate that the proposed method outperforms conventional exact, heuristic, and representative DRL-based methods in both solution quality and computational efficiency. These results verify the effectiveness of the proposed framework for dynamic resource allocation and decision-making in complex battlefield environments.

KEYWORDS

weapon-target assignment, deep reinforcement learning, spatio-temporal model, optimization methods

INTRODUCTION

In modern warfare, the strategic value of missiles is often offset by their significant resource demands [1, 2]. Adversaries increasingly adopt dispersal tactics and asymmetric strategies to improve the "salvo survival" probability of high-value assets, creating a complex targeting challenge [3-5]. This shift intensifies the difficulty of optimal resource allocation: how to effectively engage a dispersed and heterogeneous set of targets with

a limited supply of expensive advanced weapons [6-8]. The Weapon-Target Assignment (WTA) problem is a computationally complex optimization problem aimed at maximizing expected combat effectiveness while adhering to strict logistical and economic constraints [9, 10]. This research addresses this issue by developing a novel WTA model that incorporates key operational variables, with the goal of generating data-driven allocation strategies for enhanced resource utilization in dynamic and contested battlespaces.

The Weapon-Target Assignment (WTA) problem is a fundamental combinatorial optimization challenge in military operations research, focusing on the optimal allocation of a finite set of ranging weapons to engage a set of targets under multiple operational constraints [11-13]. These constraints typically include weapon-target compatibility, probabilistic destruction assessments, resource costs, and the strategic value of battlefield gains [14, 15]. Conventional research predominantly formulates WTA as a static, one-shot optimization problem (SWTA), where weapons are assigned based on a single snapshot of the battlefield to maximize an immediate utility function [16, 17]. However, this static formulation critically overlooks the inherently dynamic and uncertain nature of modern warfare. The battlefield is a complex, rapidly evolving system where adversaries are actively engaged in countermeasures, such as deploying active defense systems around high-value assets [18, 19]. These defensive actions, often unobservable in early engagement, continuously alter the probability of a missile successfully neutralizing its target. Prevailing models that assume fixed and known destruction probabilities thus embody a significant representation that severely limits practical limitations and real-world applicability [17, 20]. To bridge this gap, this study investigates a dynamic WTA (DWTA) that explicitly accounts for the temporal evolution of engagement conditions. The objective is to develop an allocation strategy that is robust to the battlefield, ultimately reducing operational costs and enhancing mission success rates in realistic combat scenarios.

Solution methods for the weapon target assignment problem are commonly categorized into exact and heuristic methods [21-23]. Exact methods have the advantage of delivering provably optimal solutions by exhaustively searching the solution space [24, 25]. For example, Haywood formulated a bilevel programming model grounded in the Stackelberg game framework, developing a mixed-integer nonlinear program to address the interceptor missile deployment problem [26]. Their approach leveraged multiple exact solvers to obtain optimal deployment strategies for interceptor missiles. To address the weapon assignment problem, Hocaoğlu formulated a nonlinear multi-objective programming framework that explicitly minimizes the survival probability of enemy assets while optimizing resource allocation [3]. Kline conducted a systematic

review of the weapon-target assignment (WTA) problem [27], offering a critical examination of the exact optimization method, and meticulously delineated the methodological strengths and practical applicability of these approaches under various operational constraints. However, exact methods become computationally intractable when applied to large-scale problems. The computational complexity of these methods grows exponentially with problem size, a fundamental limitation that prevents their use in time-sensitive operational planning for complex battlefields.

Heuristic algorithms, such as Ant Colony Optimization and Genetic Algorithms, have been extensively used to solve complex combinatorial optimization problems [28-30]. These methods utilize population-based search strategies to efficiently explore the solution space, striking a favorable balance between computational feasibility and solution quality within practical timeframes [31, 32]. As a result, they have been widely applied across various domains. Wang proposed a novel state-based modified artificial bee colony algorithm to solve the multistage weapon-target assignment (MWTA) problem [33], demonstrating superior performance in both solution quality and computational efficacy compared to existing methods. Though proving their advantages of solving optimization problems, heuristic methods inherit several limitations: they often converge to locally optimal solutions, lack theoretical guarantees of global optimality, and are highly sensitive to parameter tuning. These limitations become especially pronounced in dynamic and non-stationary environments, such as real-time battlefield scenarios. Conventional heuristic approaches primarily operate under deterministic and static assumptions, relying on iterative strategy improvements to enhance solution quality [32, 34]. However, in highly volatile combat scenarios characterized by continuously evolving threats and countermeasures, precomputed solutions quickly become obsolete. Although heuristics can be adapted through online re-optimization, this process typically requires frequent strategy adjustments and solution repairs, which significantly reduce computational efficiency. Consequently, traditional heuristic frameworks face fundamental challenges in maintaining both responsiveness and solution optimality when confronted with real-time environmental changes, thereby limiting their practical utility in dynamic, mission-critical operations.

The application of deep reinforcement learning to the WTA problem addresses key limitations of traditional methods [12, 35]. The deep reinforcement learning framework formulates WTA as a Markov Decision Process (MDP), where a deep neural network represents the policy function that maps states to assignment decisions [36, 37]. The lack of labeled training data precludes the use of supervised learning for the WTA task. Instead, deep reinforcement learning leverages a reward signal, designed to reflect mission objectives, along with

advanced training algorithms (e.g., value-based or policy-based methods) [38-40]. These algorithms train the policy network by enabling the agent to interact with a simulated combat environment, progressively refining its strategy to maximize the expected total reward [41]. This approach allows the agent to learn a near-optimal assignment policy even in high-dimensional, complex scenarios with multiple constraints. For example, Li introduced a novel hierarchical deep reinforcement learning architecture to address the sequential decision-making problem in air-to-air combat [42]. This study is specifically designed to operate in dynamic environments involving coordinated aircraft and missile engagements, aiming to derive optimal combat strategies. Zhang proposed a novel deep Q-network (DQN) architecture for air combat maneuvering decision-making [43], which was specifically designed by incorporating the principle of situational continuity. To enhance the model's capability in processing sequential information, they integrated a long short-term memory (LSTM) module, which effectively captures temporal dependencies and thereby improves both learning efficiency and overall performance.

Though significant advancements have been made in WTA research, several critical challenges remain unaddressed. First, conventional WTA studies predominantly formulate the problem in static and deterministic settings, where optimization algorithms, such as exact and heuristic, allocate resources based on a fixed battlefield snapshot. Such approaches cannot inherently adapt to dynamic changes in real-time operational environments, which characterize modern warfare. Consequently, these methods fail to maintain solution optimality under evolving tactical conditions, limiting their practical applicability. Second, although recent deep reinforcement learning methods have demonstrated promising performance in real-time multi-target attack scenarios, most existing frameworks overlook the spatio-temporal dynamics inherent in battlefield environments. In reality, the positions, statuses, and contextual relationships among missiles and targets continuously evolve. Ignoring these dynamics not only limits an agent's ability to generalize across diverse engagement scenarios but also diminishes its exploratory effectiveness during training.

To address the limitations of existing approaches, this paper proposes a novel deep reinforcement learning framework with a model (DRLSTM) framework for DWTA in complex battlefield environments. The proposed framework consists of core components: a spatial embedding and a temporal embedding. The contributions of this study can be summarized as follows:

(1) This study developed a graph convolutional network, which develops a spatial encoding to explicitly model the dynamics of the capture space. Based on a graph convolutional network, this study leverages the

topological relationships and interactions between missiles and targets at any given timestep, generating a representation.

(2) This study employs gated recurrent unit networks to capture the spatio-temporal features. This module aggregates historical state information, allowing the model to reason, enabling the trajectory and intent of entities, trajectories, and intents of entities to be represented in a single static snapshot. The synthesized spatio-temporal embedding provides a comprehensive state representation, enabling the policy network to generate context-aware missile-target assignments that adapt to the evolving tactical situation.

(3) To ensure rigorous evaluation, a high-fidelity AFSIM is utilized to generate large-scale, diverse datasets encompassing various engagement scenarios. This study conducted comparative experiments using other methods with DRLSTM. This multi-faceted comparison is designed to quantitatively validate the terms that validate solution quality, computational efficiency, robustness, and scalability.

The remainder of this paper is organized as follows. Section 2 describes the missile-target assignment problem and the fundamentals of deep reinforcement learning. Section 3 describes a novel deep reinforcement learning method. Section 4 describes the experimental results, including comparison analysis, robustness analysis, and ablation analysis. Section 5 presents the conclusions of this study.

PRELIMINARY

This section provides a comprehensive exposition of the missile-target assignment (WTA) problem by formulating it as a constrained optimization model and subsequently framing it within a reinforcement learning paradigm. First, a detailed mathematical model is established, including the definition of the objective function, system constraints, and key parameters governing the engagement scenario. Next, the problem is abstracted into a Markov Decision Process (MDP) framework, which serves as the foundational mathematical structure for applying reinforcement learning algorithms. The MDP formulation rigorously defines state space, action space, reward mechanism, and state transition dynamics, thereby creating a principled framework for developing and training intelligent decision-making agents to address the WTA problem.

Problem Description

This study investigates a multi-missile engagement scenario involving a set of targets, where the primary operational challenge is to assign each missile to an appropriate target [44, 45]. The assignment strategy is a critical factor in determining the overall combat effectiveness of the missile group. Therefore, the main

objective of this research is to maximize the combat benefit, which includes both total combat effectiveness and combat cost. This objective can be formally expressed as follows:

$$\max F(x) = T(x) - U(x) \tag{1}$$

Where $F(x)$ denotes the benefit of combat. $T(x)$ and $U(x)$ respectively denote combat effectiveness and combat cost, and their formulas are defined as follows:

$$maxT(x) = \sum_{j=1}^{N} \left[1 - \prod_{i} (1 - p_{ij})^{x_{ij}} \right] v_j \quad \forall i \in M \tag{2}$$

Where formula (2) denotes combat effectiveness. v_j denotes the destruction value. p_{ij} ($0 \leq p_{ij} \leq 1$) denotes the destruction probability of i with respect to j . $1 - \prod_i (1 - p_{ij})^{x_{ij}}$ denotes the joint destruction probability of multiple missiles. In practical scenarios, the probability of a missile destroying a target is a complex function influenced by factors such as distance, angle of attack, and defensive capabilities. To simplify this process, this study employs a Gaussian function to randomly generate penetration values representing the probability of destruction.

$$minU(x) = \sum_{i=1}^{M} \sum_{j=1}^{N} u_i x_{ij} \tag{3}$$

Where formula (3) denotes the combat cost. u_i denotes the missile cost of i . In formulas (2) and (3), x_{ij} denotes target j of missile i . M and N denote the number of missiles and targets, respectively.

Markov Decision Process

Deep reinforcement learning is fundamentally based on the theory of Markov Decision Process (MDP), which provides the standard framework for modeling goal-directed learning in interactive environments [46, 47]. An MDP is defined by the quintuple (S, A, P, R, γ) , where S is a finite set of states describing the environment; A is a finite set of actions available to the agent; P is the state transition probability function; R is the immediate reward function; and γ is a discount factor that determines the importance of future rewards. This formalism establishes the foundational assumptions and components that deep reinforcement learning algorithms are designed to leverage and optimize, especially when state and action spaces are too large

for traditional tabular methods, thereby necessitating the use of deep neural networks for function approximation. Additionally, the WTA process is formulated under the assumption that the decision-making agent receives fused battlefield information from the sensing and command system. Accordingly, the environment is modeled as fully observable at the decision level. This assumption simplifies the analysis and allows us to focus on the dynamic assignment mechanism. Nevertheless, in real combat scenarios, battlefield information is often incomplete, delayed, or noisy. Therefore, the current formulation does not fully account for the fog of war and information uncertainty.

The application of deep reinforcement learning to complex decision-making problems, in missile-target assignment, requires a rigorous mathematical foundation. This foundation is provided by the MDP framework, which formalizes the interaction between an intelligent agent and its environment as a sequential decision-making problem under uncertainty. The successful formulation of an MDP is a critical prerequisite for the effective application of any deep reinforcement learning algorithm. The MDP is designed as follows: $s_t \in S$ is the state of the environment at time step t . The state $s_t = [o_t, h_t, j_t]$ is a composite vector that integrates. The g_t is the radius of destruction for all missiles, the h_t is the circular probability error of all missiles, j_t is the value of all targets. $a_t \in A$ denotes the action taken at time step t . The A is determined by the granularity of control. P denotes the probability of transitioning to the state s' upon acting a in state s . In this study, the probability of a target being destroyed is given an assignment. R is the objective function (formula 1). γ is set to 1 in this study.

METHOD

To address the challenges outlined above, this study presents a deep reinforcement learning framework based on a spatio-temporal model for the real-time optimization of missile-target assignments. As shown in Figure 1, the proposed method extracts both static features, such as the inherent attributes of missiles and targets, and dynamic features, including the evolving spatial relationships between missiles and targets over time. Initially, linear encoders process missile and target data separately. These encoded inputs are then passed through spatial and temporal encoders to dynamically capture spatio-temporal interactions at each decision step. A decoder subsequently generates optimal missile-target pairings to maximize overall destruction efficiency. To capture the rapid kinematic variations of high-speed weapons, the temporal sampling interval is set to 0.2 s. The model is trained end-to-end using an actor-critic algorithm, enabling the development of an effective policy function for real-time assignment.

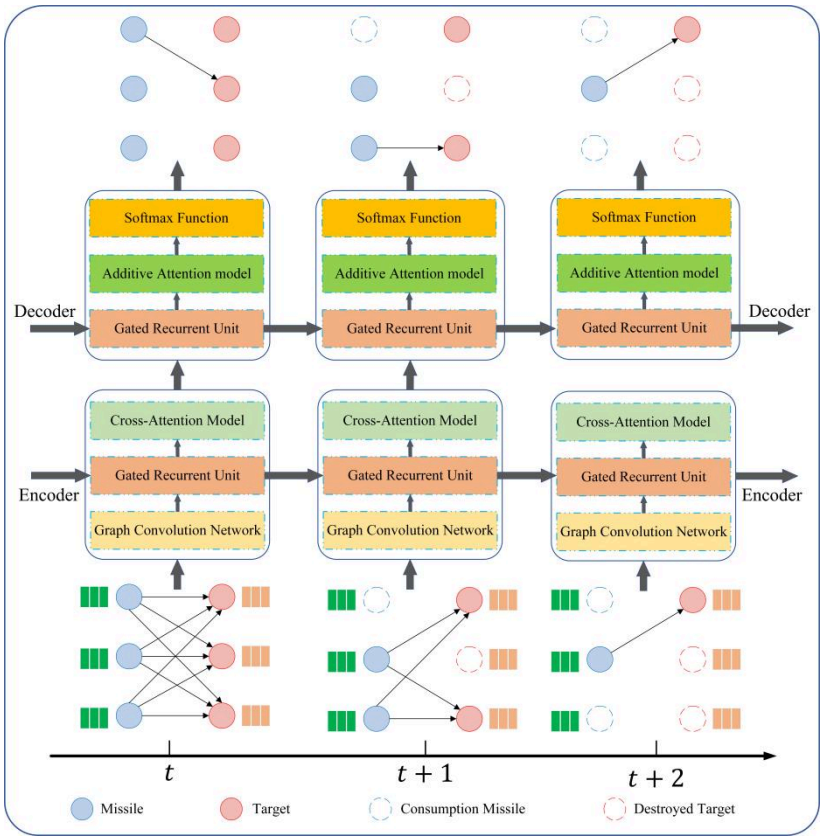

Figure 1. Total Workflow of Our DRLSTM

The encoder is designed to learn a unified spatiotemporal context for the engagement scenario. It operates in two distinct stages: spatial encoding and temporal encoding. The spatial encoder utilizes graph models to capture the topological relationships and interactions among missiles and targets at each snapshot in each step time. Subsequently, the temporal encoder processes the sequence of these spatial snapshots using recurrent or self-attention-based models. Specifically, the features of missiles and targets are first extracted at each time step. The missile's features include the geographic information of missile $X(x_0, x_1, ..., x_n)$, radius of destruction of the missile $O_t(o_0, o_1, ..., o_t)$, and circular probability error of the missile $H_t(h_0, h_1, ..., h_t)$. The X , O_t , and H_t are concatenated $\dot{L}_t = [X, O_t, H_t]$. The location information $\ddot{X}(\ddot{x}_0, \ddot{x}_1, ..., \ddot{x}_n)$ and value information $J_t(j_0, j_1, ..., j_t)$ in the targets are concatenated $\ddot{L}_t = [\ddot{X}, J_t]$. Then, $\dot{L}_t$ and $\ddot{L}_t$ are modeled as follows:

$$\dot{F}_t = \dot{L}_t W_0 + B_0 \tag{4}$$

$$\ddot{F}_t = \ddot{L}_t W_1 + B_1 \tag{5}$$

Where W_0 , W_1 , B_0 , and B_1 are learnable parameters. To obtain comprehensive features of missiles and targets, $\dot{F}_t$ and $\ddot{F}_t$ are concatenated $\dddot{F}_t = [\dot{F}_t, \ddot{F}_t]$.

The spatial structure of missiles and targets changes over time. The adjacent matrix $A^t = (a^t_{0\bullet 1}, a^t_{0\bullet 2}, ..., a^t_{i\bullet j})$ is expressed spatial information at each time step. The $a^t_{i\bullet j}$ is defined as follows:

$$a^t_{i\bullet j} = \begin{cases} 1, i \ \text{ and } \ j \ \text{ are adjacent} \\ 0, \text{other} \end{cases} \tag{6}$$

Where $a^t_{i\bullet j}$ denotes the spatial information between missiles and targets at each time step. $a^t_{i\bullet j} = 1$ denotes that j is the target of missile i . Otherwise, missile i cannot attack target j .

Then, the graph convolutional network (GCN) is a spatial encoder, which extracts spatial missiles and targets [48, 49]. The relevant definitions are as follows:

$$G_t = \sigma(\hat{A}^t \ \text{Relu} \ (\hat{A}^t \dddot{F}_t W_3) W_4) \tag{7}$$

Where W_3 and W_4 are learnable parameters. $\hat{A}^t = \tilde{D}^{-1/2}(A^t + I_N)\tilde{D}^{-1/2}$ is the preprocessing step, I_N is the self-connections matrix. $\sigma(\bullet)$ represents the sigmoid function for a nonlinear function.

The Gated Recurrent Unit (GRU) is employed as a temporal encoder to model the evolution of the missile–target engagement state and to extract sequential spatio-temporal features [50, 51]. In this study, each time step in the GRU input sequence corresponds to a fixed sampling interval of $\Delta t =$ seconds, which is consistent with the update frequency of the data source. Specifically, the time step Δt was set to 0.2 s. This choice satisfies the Nyquist sampling criterion for targets with velocities up to 250 m/s, ensuring that rapid kinematic changes in high-speed targets (e.g., missiles) are adequately captured. The relevant definitions are as follows:

$$Z_t, h_t = \text{GRU} \ (G_{t-1}, h_{t-1}) \tag{8}$$

Finally, this study employs cross-attention to the spatio-temporal dependencies between spatio-temporal features and features of missiles. The relevant definitions are as follows:

$$\hat{F}_t = \text{Cross-Attention} \ (Z_t, \dot{F}_t) \tag{9}$$

Where $\dot{F}_t$ is the query and value, Z^t is the key.

Decoder

The decoder module is designed to sequentially generate missile-target assignment pairs at each time step. This iterative process is governed by a GRU model, which aggregates historical assignment decisions and the evolving states of targets to form a dynamic, contextual query vector. This query, representing the current decision context, is then fed into a multi-head attention mechanism. The attention model computes compatibility scores between this query and the encoded features of all available missiles, effectively identifying the most suitable missile-target pair based on the current tactical context. The detailed operational workflow of the decoder is outlined as follows:

$$I_t, \ddot{h}_t = \mathrm{GRU}\ (\ddot{F}_{t-1}, \ddot{h}_{t-1}) \tag{10}$$

$$\check{F}_t = W_2 \tanh(W_3 I_t + W_4 \hat{F}_t) \tag{11}$$

$$P^t = \mathrm{softmax}\ (\check{F}_t) \tag{12}$$

Where W_2 , W_3 , and W_4 are learnable parameters. The formula (11) is an additive attention model. P^t is the probability of candidate missiles and targets in the formula (12).

Training Method

To optimize the policy parameters [52, 53], this approach utilizes two neural networks: a parameterized actor, which defines the policy by outputting action probabilities for weapon-target pairing, and a critic, which approximates the value function to evaluate the quality of states visited under the actor's policy. Although the two networks may share a common feature extraction backbone, their output layers and optimization objectives remain distinct. The critic's evaluation is used to formulate the policy gradient, guiding the actor's updates toward regions of higher expected return. The resulting loss function for our deep reinforcement learning agent is defined as follows:

$$J_\pi(\theta) = E[(R(\pi) - B(S)) \log p(a^t \mid s^t, \theta)] \tag{13}$$

Where $R(\pi)$ is a cumulative reward in formula (13), $B(S)$ is a critic's function.

Algorithm 1. The workflow actor-critic algorithm.

Training method: actor-critic algorithm
Input: policy model, training set S , training epoch N , batch size B , customers M
Output: parameter θ
Initialize parameters θ of policy model
for each = 1, 2, …, N **do**
for batch = 1, 2, …, B **do**
Generate training set S
while $m < M$ **do**
Compute encoding features by the encoder
Compute the decoding information by the decoder
Compute the missile and target
end
Calculate the cumulative rewards $R(\pi)$
Generate the loss function: $J_{\pi}(\theta) = E[(R(\pi) - B(S)) \log p(a^t \mid s^t, \theta)]$
Update the parameters of the policy model
end
end

EXPERIMENTS

This section presents a comprehensive empirical evaluation of the proposed strategy model through simulations. A high-fidelity simulation environment is established to generate the experimental data necessary for performance assessment. The evaluation is organized into three distinct phases: (1) benchmarking the model against classical optimization algorithms to quantitatively validate its solution quality and computational efficiency, with a primary focus on the optimality gap. (2) The rigorous assessment of stability involves various operational uncertainties in the battlefield environment; (3) To verify the contribution of each core component, ablation studies are conducted, which involve altering individual modules to isolate their impact on overall performance.

Experimental data

This study utilizes the Advanced Framework for Simulation, Integration, and Modeling (AFSIM) to generate a comprehensive dataset for model training and validation. The simulated battlespace includes two heterogeneous missile types: M1 (Cost = 1.0, P = 0.35) and M2 (Cost = 1.25, P = 0.70), which engage four classes of targets (T1, T2, T3, T4) with respective values of 1, 2, 3, and 8. And this study generated four types of missile-target experimental groups: (12,6), (24,12), (36,18),(72,24), and (96,48). In each group, a total of 100,000

stochastic engagement scenarios were generated, with 70,000 instances allocated for training and 30,000 for validation. The time step Δt was set to 0.2 s. All algorithms implemented in Python 3.8 are executed on a high-performance workstation equipped with NVIDIA RTX 3090 GPUs to ensure efficient training and inference.

Baseline Algorithms

To rigorously evaluate the performance of the proposed methodology, a comprehensive benchmarking framework was established, encompassing three distinct categories of comparative approaches: commercial mathematical optimization solvers, heuristic algorithms, and state-of-the-art deep reinforcement learning methods.

Commercial Solver

The commercial solver Gurobi, a widely recognized mathematical programming optimizer, is used to compute benchmark optimal solutions [54]. To balance solution optimality with computational tractability, Gurobi is applied exclusively to a small-scale scenario featuring a (12, 6) missile-target configuration in 120 seconds.

Heuristic methods

Genetic Algorithm (GA) and Ant Colony Optimization (ACO) are classical optimization techniques [55,56]. As a population-based evolutionary algorithm, GA is employed for its proven ability to efficiently explore complex solution spaces and generate high-quality, near-optimal solutions for combinatorial optimization problems. Inspired by the foraging behavior of real ant colonies, ACO is a swarm intelligence algorithm widely applied in academic research to solve path-finding and assignment problems, making it a relevant benchmark for this study.

Deep Reinforcement Learning

The proposed model is further compared against contemporary deep reinforcement learning approaches to assess its relative advancement. A standard deep reinforcement learning baseline incorporating Recurrent Neural Networks and attention mechanisms is commonly applied across various sequential decision-making and combinatorial optimization tasks [57]. A state-of-the-art approach predicated based on the Transformer architecture is currently at the forefront of research for solving complex optimization problems due to its superior sequence modeling and ability to capture dependencies [58].

Comparison Analysis

Table 1 summarizes the experimental results across five distinct scenarios. The “Objective” column indicates the average combat benefit achieved in each case, while the “Gap” column quantifies the performance difference relative to the Genetic Algorithm (GA) baseline. The best optimization results are indicated in bold.

Table 1. Experimental results in multiple methods

Method	(12, 6) Objective	(12, 6) Gap/%	(24, 12) Objective	(24, 12) Gap/%	(36, 18) Objective	(36, 18) Gap/%	(48, 24) Objective	(48, 24) Gap/%	(96, 48) Objective	(96, 48) Gap/%
Gurobi	35.96	6.81	/	/	/	/	/	/	/	/
GA	33.67	/	56.77	/	83.01	/	110.23	/	168.08	/
ACO	34.01	1.01	59.39	4.62	87.44	5.34	114.76	4.11	176.92	5.26
DRL	34.75	3.21	60.86	7.21	88.86	7.05	119.86	8.74	183.07	8.92
DRLTSF	35.11	4.28	60.73	6.98	89.00	7.22	121.54	10.26	189.49	12.74
DRLSTM	**35.87**	**6.53**	**61.55**	**8.42**	**90.71**	**9.26**	**124.27**	**12.74**	**193.46**	**15.10**

The experimental results summarized in Table 2 and Table 3 demonstrate the superior performance of the proposed method. Although the commercial solver Gurobi attains a slightly better objective value, it requires a computationally intensive runtime of 120 seconds. In contrast, the proposed DRLSTM method, as a generative model, delivers near-optimal solutions within milliseconds, highlighting a significant advantage in computational efficiency for time-sensitive applications. When benchmarked against the Genetic Algorithm (GA), the proposed method demonstrates a significant performance improvement ranging from 6.53% to 15.10%. This performance gap widens as the scene scale increases, highlighting the superior scalability of our approach. Furthermore, the proposed method consistently outperforms existing deep reinforcement learning baselines, underscoring the benefits of incorporating dynamic spatiotemporal features into the optimization model. Collectively, these findings validate the proposed method’s balance between high solution quality and real-time capability, demonstrating the significant potential of deep reinforcement learning for complex mission planning.

Table 2. Computational time of different methods

Method	(12,6)	(24,12)	(48,24)	(96,48)
Gurobi	1.20e+10 s	/	/	/
GA	2.22e-2 s	4.37e-2 s	2.79e-1 s	1.83e+0 s
ACO	1.94e+0 s	4.62e+0 s	1.57e+1 s	3.23e+1 s
DRL	4.57e-4 s	2.69e-3 s	5.37e-3 s	4.19e-2 s
DRLTSF	8.28e-4 s	7.38e-3 s	1.20e-2 s	9.38e-2 s
DRLSTM	2.18e-3 s	1.45e-2 s	4.82e-2 s	1.14e-1 s

Robustness Analysis

Beyond mere optimality under ideal conditions, the stability and robustness of a model are paramount in assessing its practical viability and generalization capability, especially for mission-critical applications. Model stability is defined as the consistency and repeatability of performance under nominal operating conditions, ensuring that the model's output does not exhibit significant variance due to intrinsic stochastic factors such as parameter initialization or data sampling order. Conversely, model robustness refers to the capacity to maintain core functional performance and degrade gracefully when subjected to external perturbations, including input noise, environmental dynamics, and distributional shifts, rather than failing catastrophically.

To quantitatively evaluate these indicators, we conducted a comprehensive empirical analysis across four distinct operational scenarios. As shown in Figure 2, the performance distribution of each model, including the proposed DRLSTM and several benchmark algorithms, was visualized using box plots. These plots effectively summarize key statistical measures, such as the mean and standard deviation of the performance metrics. The experimental results demonstrate that the proposed DRLSTM method produces box plots that are positioned lower (indicating superior average performance) and are shorter in length (signifying lower variance and greater stability) compared to all alternative approaches. Specifically, methods based on deep reinforcement learning consistently outperformed traditional heuristic algorithms in both solution quality and reliability in the (24, 12) engagement scenario. Furthermore, among the DRL methods, DRLSTM demonstrated clear superiority over its counterparts, such as DRL and DRLTSF. This performance advantage became increasingly pronounced as the scenario complexity escalated with larger numbers of missiles and targets. The consistent outperformance of DRLSTM, characterized by higher median performance and narrower interquartile ranges across all tested conditions, provides compelling evidence of its enhanced stability and robustness, underscoring its suitability for real-world deployment where uncertainty and dynamic changes are inherent. Therefore, the comprehensive experiments on stability and robustness verify that our method is not only

effective in ideal conditions but also reliable and resilient in the face of uncertainties and adversarial situations, which is crucial for real-world deployment.

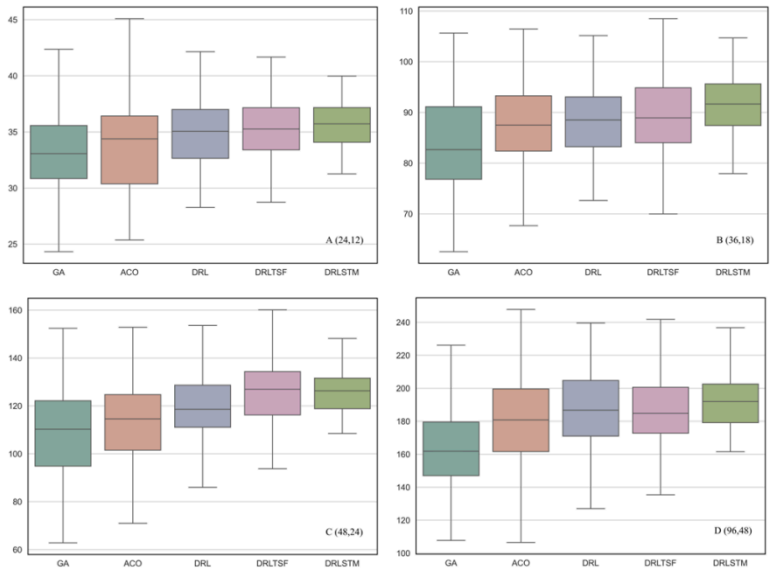

Figure 2. The Box Plots of Different Methods

Ablation Analysis

To quantitatively assess the contribution of spatio-temporal features to the optimization performance of the proposed DRLSTM, controlled ablation studies were conducted. Specifically, we developed two simplified variants: the Deep Reinforcement Learning Temporal Model (DRLTM), created by removing the spatial module, and the Deep Reinforcement Learning Spatial Model (DRLSM), created by removing the temporal module. The evolutionary trends of the average optimization results for these ablated models, alongside the complete DRLSTM, are compared in Figure 3 under the (24, 12) and (36, 18) scenarios. The experimental results clearly demonstrate that both DRLTM and DRLSM exhibit statistically significant performance degradation compared to DRLSTM. The final performance of both ablated models is suboptimal, and their convergence speeds are notably slower. These findings collectively confirm that omitting either spatial or temporal features results in an incomplete state representation, causing the model to overlook critical task information and ultimately leading to poorer optimization performance and convergence behavior. This ablation analysis conclusively demonstrates the necessity of integrating both spatial and temporal modules within the proposed architecture.

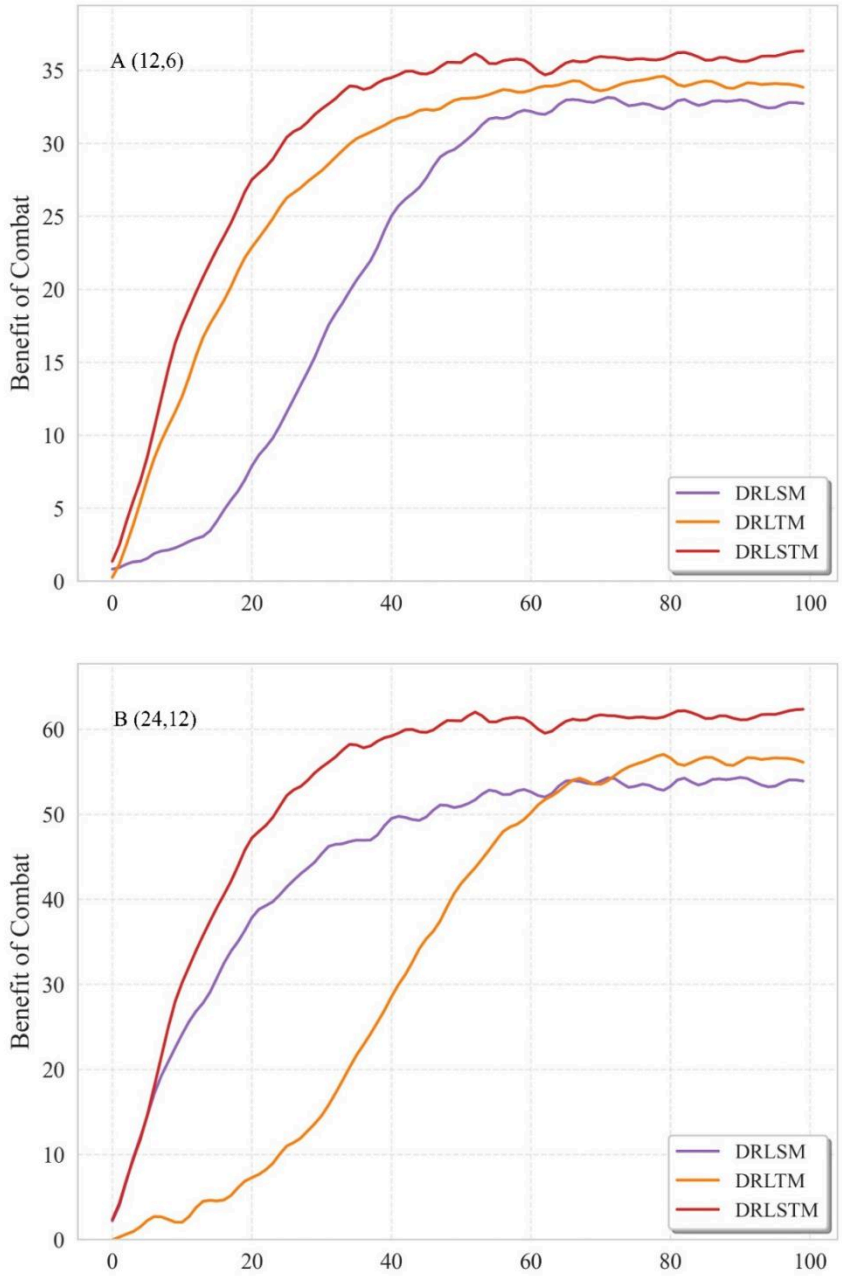

Figure 3. Convergence Curve of Ablation experiments

CONCLUSION

WTA is a classic NP-hard combinatorial optimization problem in modern command and control systems. Optimizing WTA efficiency is crucial, as it directly minimizes operational costs and maximizes the engagement effectiveness of weapon systems. However, the inherent complexity of this problem is significantly amplified by the dynamic and uncertain nature of the modern battlefield, where real-time information updates demand highly adaptive decision-making. To address this critical challenge, this paper introduces a novel deep rein-

forcement learning framework that dynamically encodes the spatio-temporal features of the engagement environment. The proposed method is empirically demonstrated to outperform established commercial solvers, classical metaheuristics, and other state-of-the-art DRL baselines. By leveraging real-time situational awareness, our model facilitates more intelligent resource allocation, thereby significantly enhancing weapon system utilization and reducing overall operational costs.

Despite these promising results, our approach has several limitations that open avenues for future research. First, the current DRLSTM framework is primarily suitable for small- to medium-scale MTA problems. As the dimensionality of the battlespace increases, the computational cost of training rises nonlinearly, and the policy's decision-making efficiency may decline. Therefore, developing scalable DRL architectures and efficient training paradigms for large-scale and ultra-large-scale scenarios is a primary objective for future work. Second, a limitation of the present study is that the graph model adopts homogeneous node processing and does not explicitly account for heterogeneous target categories, such as decoy-capable, heavily armored, or highly maneuverable threats. Although the proposed framework performs effectively under the considered settings, extending it with heterogeneous graph learning or threat-specific representation mechanisms would further improve its realism and generalization capability in complex combat scenarios.

Author Contributions

Conceptualization – Surname X, Surname Y and Surname Z; methodology – Surname X and Surname Y; formal analysis – Surname X and Surname Y; investigation – Surname X; resources – Surname X; writing-original draft preparation – Surname X, Surname Y and Surname Z; writing-review and editing – Surname X and Surname Y; visualization – Surname X; supervision – Surname X. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgements

Not applicable.

REFERENCES

[1] LUO H, JIANG R, WANG G. Multi-source information fusion based on evidence reasoning using double frames of discernment for estimating the number of remaining missiles. Expert Systems with Applications. 2025; 260: 125373. doi: 10.1016/j.eswa.2024.125373

[2] SALIH A, MOSHAIOV A. Neuro-evolution-based generic missile guidance law for many-scenarios. Applied Soft Computing. 2024; 152: 111210. doi: 10.1016/j.asoc.2023.111210

[3] HOCAOĞLU M F. Weapon target assignment optimization for land based multi-air defense systems: A goal programming approach. Computers & Industrial Engineering. 2019; 128: 681-689. doi: 10.1016/j.cie.2019.01.015

[4] TASHAKORI S, RANJBAR M, BALOCHIAN S, et al. Dynamic soft-kill weapon-target assignment in naval environments. Computers & Industrial Engineering. 2024; 197: 110606. doi: 10.1016/j.cie.2024.110606

[5] TUNCER O, CIRPAN H A. Adaptive fuzzy based threat evaluation method for air and missile defense systems. Information Sciences. 2023; 643: 119191. doi: 10.1016/j.ins.2023.119191

[6] LEE H, CHOI B J, KIM C O, et al. Threat evaluation of enemy air fighters via neural network-based Markov chain modeling. Knowledge-Based Systems. 2017; 116: 49-57. doi: 10.1016/j.knosys.2016.10.032

[7] SUMMERS D S, ROBBINS M J, LUNDAY B J. An approximate dynamic programming approach for comparing firing policies in a networked air defense environment. Computers & Operations Research. 2020; 117: 104890. doi: 10.1016/j.cor.2020.104890

[8] WANG D, XIN B, WANG Y, et al. Constraint-Feature-Guided Evolutionary Algorithms for Multi-Objective Multi-Stage Weapon-Target Assignment Problems. Journal of Systems Science and Complexity. 2025; 38(3): 972-999. doi: 10.1007/s11424-025-4232-2

[9] XU Q Q, LI K Q, YUE Z Q, et al. Weapon Target Allocation Based on GA-APSO Algorithm. IEEE Access. 2024; 12: 164337-164351. doi: 10.1109/ACCESS.2024.3491773

[10] SHEN Z S, LIU T, MA L. Dynamic weapon target assignment of USV based on hybrid compact Genetic algorithm. In: Proceedings of the 2023 35th Chinese Control and Decision Conference (CCDC). 2023. doi: 10.1109/CCDC58219.2023.10327049

[11] FATIMA R, SADIQ R, ULLAH I, et al. Multiple passive-sensor distributed target tracking approach with Machine Learning Feedback. Expert Systems with Applications. 2024; 238: 122344. doi: 10.1016/j.eswa.2023.122344

[12] XIN B, CHEN J, ZHANG J, et al. Efficient Decision Makings for Dynamic Weapon-Target Assignment by Virtual Permutation and Tabu Search Heuristics. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2010; 40(6): 649-662. doi: 10.1109/TSMCC.2010.2049261

[13] HUAIPING C, JINGXU L, YINGWU C, et al. Survey of the research on dynamic weapon-target assignment problem. Journal of Systems Engineering and Electronics. 2006; 17(3): 559-565. doi: 10.1016/S1004-4132(06)60097-2

[14] SUN Z, PIAO H, YANG Z, et al. Multi-agent hierarchical policy gradient for Air Combat Tactics emergence via self-play. Engineering Applications of Artificial Intelligence. 2021; 98: 104112. doi: 10.1016/j.engappai.2020.104112

[15] CHEN L, QI G, LI Y, et al. Optimal Control Strategies in Multi-Pursuit-Multi-Evasion Differential Games with Communication Graphs. In: Proceedings of the 2025 4th Conference on Fully Actuated System Theory and Applications (FASTA). 2025. doi: 10.1109/FASTA65681.2025.11138158

[16] LI J, WU G, WANG L. A comprehensive survey of weapon target assignment problem: Model, algorithm, and application. Engineering Applications of Artificial Intelligence. 2024; 137: 109212. doi: 10.1016/j.engappai.2024.109212

[17] LILES J M, ROBBINS M J, LUNDAY B J. Improving defensive air battle management by solving a stochastic dynamic assignment problem via approximate dynamic programming. European Journal of Operational Research. 2023; 305(3): 1435-1449. doi: 10.1016/j.ejor.2022.06.031

[18] CHEN Y, LUO H, WANG G. Expert Experience Soft Actor-Critic for Unmanned Aerial Vehicle Dynamic Target Assignment. J Aerosp Inf Syst. 2025; 22: 379-390. doi: 10.2514/1.i011435

[19] HUGHES M S, LUNDAY B J. The Weapon Target Assignment Problem: Rational Inference of Adversary Target Utility Valuations from Observed Solutions. Omega. 2022; 107: 102562. doi: 10.1016/j.omega.2021.102562

[20] ZHAO F, YANG T, XU T, et al. A multi-objective double Q-learning-based hyper-heuristic algorithm for aluminum production and transportation integrated scheduling problem. Engineering Applications of Artificial Intelligence. 2025; 161: 112169. doi: 10.1016/j.engappai.2025.112169

[21] ABED-ALGUNI B H. EvoMapX: An explainable framework for metaheuristic optimization algorithms. Expert Systems with Applications. 2026; 298: 129514. doi: 10.1016/j.eswa.2025.129514

[22] LI W, DU G, YUE X. A multi-time-window multi-objective hybrid fleet home health care routing optimization problem considering caregiver utilization and compatibility. Computers & Operations Research. 2026; 185: 107288. doi: 10.1016/j.cor.2025.107288

[23] MAMAGHANI E J, BATTAIA O. Toward sustainable and customer-centric reverse logistics: Machine learning-enhanced multi-objective optimization. Computers & Industrial Engineering. 2026; 211: 111622. doi: 10.1016/j.cie.2025.111622

[24] WANG T, WU D, YANG J. Efficiency with consent: Permutable queueing in on-demand services. Omega. 2026; 138: 103361. doi: 10.1016/j.omega.2025.103361

[25] ZHAO S, LIU J, KADZIŃSKI M, et al. A Probabilistic preference learning approach for multiple criteria ranking in dynamic decision context. European Journal of Operational Research. 2026; 330(2): 558-574. doi: 10.1016/j.ejor.2025.08.008

[26] HAYWOOD A B, LUNDAY B J, ROBBINS M J. Intruder detection and interdiction modeling: A bilevel programming approach for ballistic missile defense asset location. Omega. 2022; 110: 102640. doi: 10.1016/j.omega.2022.102640

[27] KLINE A, AHNER D, HILL R. The Weapon-Target Assignment Problem. Computers & Operations Research. 2019; 105: 226-236. doi: 10.1016/j.cor.2018.10.015

[28] XIN B, CHEN J, PENG Z, et al. An Efficient Rule-Based Constructive Heuristic to Solve Dynamic Weapon-Target Assignment Problem. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2011; 41(3): 598-606. doi: 10.1109/TSMCA.2010.2089511

[29] LI W, WANG R, HENG Y, et al. Knowledge-Guided Evolutionary Optimization for Large-Scale Air Defense Resource Allocation. IEEE Transactions on Artificial Intelligence. 2024; 5(12): 6267-6279. doi: 10.1109/TAI.2024.3375263

[30] LI X, ZHOU D, PAN Q, et al. Weapon-Target Assignment Problem by Multiobjective Evolutionary Algorithm Based on Decomposition. Complex. 2018; 2018: 19. doi: 10.1155/2018/8623051

[31] OZSOYDAN F B, GÖLCÜK İ, DURMAZ E D. A hyper-heuristic enhanced neuro-evolutionary algorithm with self-adaptive operators and various activation functions for classification problems. Neural Networks. 2025; 190: 107751. doi: 10.1016/j.neunet.2025.107751

[32] GUANG P, YANGWANG F, SHAOHUA C, et al. A Hybrid Multiobjective Discrete Particle Swarm Optimization Algorithm for Cooperative Air Combat DWTA. Journal of Optimization. 2017; 2017: 1-12. doi: 10.1155/2017/8063767

[33] WANG X, ZHANG Y, WANG G. Target assignment for multiple stages of weapons systems using a deep Q-learning network and a modified artificial bee colony method. Computers and Electrical Engineering. 2024; 118: 109378. doi: 10.1016/j.compeleceng.2024.109378

[34] ZHENG S, LI Z, ZHENG W, et al. A novel scaling-based landing first constructive heuristic algorithm for aircraft scheduling and parking problem in multi-runway airports. Transportation Research Part E: Logistics and Transportation Review. 2025; 204: 104389. doi: 10.1016/j.tre.2025.104389

[35] LIU J-Y, WANG G, FU Q, et al. Task assignment in ground-to-air confrontation based on multiagent deep reinforcement learning. Defence Technology. 2023; 19: 210-219. doi: 10.1016/j.dt.2022.04.001

[36] BEN ELALLID B, BENAMAR N, BAGAA M, et al. Secure and efficient vehicle control of autonomous vehicles using federated deep reinforcement learning. Applied Soft Computing. 2025; 185: 113924. doi: 10.1016/j.asoc.2025.113924

[37] CHIBOUB A, FRANCOIS J, ALIX T, et al. Contribution of deep reinforcement learning to solve reconfigurable facilities layout problems. Manufacturing Letters. 2025; 46: 16-20. doi: 10.1016/j.mfglet.2025.09.003

[38] GUO K, MA J, WU J, et al. Ai-driven energy optimization for mineral transport: A novel deep reinforcement learning approach for smart pumping station scheduling. Information Sciences. 2026; 728: 122780. doi: 10.1016/j.ins.2025.122780

[39] HUANG J, ZHOU R, LI M, et al. From black-box to white-box: Interpretable deep reinforcement learning with Kolmogorov-Arnold networks for autonomous driving. Transportation Research Part C: Emerging Technologies. 2026; 182: 105386. doi: 10.1016/j.trc.2025.105386

[40] YOUSIF M Z G, KOLESOVA P, YANG Y, et al. Optimizing flow control with deep reinforcement learning: Plasma actuator placement around a square cylinder. Physics of Fluids. 2023. doi: 10.1063/5.0174724

[41] XU L, HU B, GUAN Z, et al. Multi-agent Deep Reinforcement Learning for Pursuit-Evasion Game Scalability. In: Proceedings of 2019 Chinese Intelligent Systems Conference. 2020. doi: 10.1007/978-981-32-9682-4_69

[42] LI L, ZHANG X, QIAN C, et al. Cross coordination of behavior clone and reinforcement learning for autonomous within-visual-range air combat. Neurocomputing. 2024; 584: 127591. doi: 10.1016/j.neucom.2024.127591

[43] ZHANG J-D, YU Y-F, ZHENG L-H, et al. Situational continuity-based air combat autonomous maneuvering decision-making. Defence Technology. 2023; 29: 66-79. doi: 10.1016/j.dt.2022.08.010

[44] WANG D, XIN B, ZHANG J, et al. A Clustering-Based Adaptive Hybrid Algorithm for the Stochastic Resource Allocation Problem With Time Windows. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2025; 55(12): 9468-9482. doi: 10.1109/TSMC.2025.3614245

[45] LUO W, J L, LIU K, et al. Learning-Based Policy Optimization for Adversarial Missile-Target Assignment. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2022; 52(7): 4426-4437. doi: 10.1109/TSMC.2021.3096997

[46] LIU J, WANG M, LIU K. Interpretable deep reinforcement learning optimizes emergency response for China's freezing rain-impacted roads. Transportation Research Part D: Transport and Environment. 2026; 150: 105076. doi: 10.1016/j.trd.2025.105076

[47] NGUYEN H, THUDUMU S, DU H, et al. A comprehensive survey on deep reinforcement learning in object tracking. Machine Learning with Applications. 2025; 22: 100745. doi: 10.1016/j.mlwa.2025.100745

[48] SUN J, ZHANG Y, GUO W, et al. Neighbor Interaction Aware Graph Convolution Networks for Recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1289-1298. doi: 10.1145/3397271.3401123

[49] SHI N, CHEN H, CHEN L, et al. Chaotic dual-feature graph convolutional network (CDF-GCN) for traffic speed forecasting. Applied Soft Computing. 2026; 186: 114096. doi: 10.1016/j.asoc.2025.114096

[50] MUDUSU R, K V. A heuristic approach to HLCA-based feature extraction using a GRU model for Parkinson's disease detection. Neurocomputing. 2026; 664: 132061. doi: 10.1016/j.neucom.2025.132061

[51] QIN Z, ZHANG J, WANG W, et al. Adaptive AI-driven incentive framework using GCN-GRU and multi-objective optimization for sustainable urban behavior in smart cities. Sustainable Cities and Society. 2025; 135: 106972. doi: 10.1016/j.scs.2025.106972

[52] CHEN H, SHEN L-Y, WANG C, et al. Multi Actors-Critic based particle swarm optimization algorithm. Neurocomputing. 2025; 624: 129460. doi: 10.1016/j.neucom.2025.129460

[53] LI B, WANG H. Improving mobility management in LEO Satellite networks utilizing Soft Actor-Critic algorithms. Computer Networks. 2025; 271: 111608. doi: 10.1016/j.comnet.2025.111608

[54] RAHMANIANI R, CRAINIC T G, GENDREAU M, et al. The Benders decomposition algorithm: A literature review. European Journal of Operational Research. 2017; 259(3): 801-817. doi: 10.1016/j.ejor.2016.12.005

[55] GAO S, ZUO L, LU X, et al. Cooperative target allocation for heterogeneous agent models using a matrix-encoding genetic algorithm. Journal of Information and Intelligence. 2025; 3(2): 154-172. doi: 10.1016/j.jiixd.2024.07.002

[56] CUI J, WU L, HUANG X, et al. Multi-strategy adaptable ant colony optimization algorithm and its application in robot path planning. Knowledge-Based Systems. 2024; 288: 111459. doi: 10.1016/j.knosys.2024.111459

[57] NAZARI M, OROOJLOOY A, SNYDER L V, et al. Reinforcement Learning for Solving the Vehicle Routing Problem. In: Neural Information Processing Systems. 2018.

[58] YANG H, ZHAO M, YUAN L, et al. Memory-efficient Transformer-based network model for Traveling Salesman Problem. Neural Networks. 2023; 161: 589-597. doi: 10.1016/j.neunet.2023.02.014