Game-Theoretic and Inverse Reinforcement Learning-Based Control for Vehicle Formation Under DoS Attacks

Xiaoping Zhao, Jia Guo, Jinliang Liu, *Senior Member, IEEE*, and Wenbin Huang

***Abstract*—To address the performance degradation in vehicle formation control caused by communication disruptions under denial-of-service (DoS) attacks, this article proposes a secure control method that integrates a nonzero-sum game and inverse reinforcement learning (IRL). At the game-theoretic layer, a dynamic game model accounting for the attacker's energy cost is constructed to capture the adversarial interaction and strategy spaces between DoS attacks and the formation controller, with an approximate Nash equilibrium solution derived for both attack and defense strategies. At the reward learning layer, an IRL approach is introduced to adaptively learn the weight parameters of multiobjective performance indices from offline demonstration data, thereby mitigating the sensitivity of system performance to manual weight tuning. At the control implementation layer, based on the learned reward weights, a neural network (NN) is employed to approximate the solution of the Hamilton–Jacobi (HJ) equation, thereby constructing a computable feedback control law. Simulation results demonstrate that the proposed method can effectively suppress formation tracking errors in DoS attack scenarios and exhibits superior robustness, adaptability, and energy efficiency compared to traditional methods.**

***Index Terms*—Denial-of-service (DoS) attacks, inverse reinforcement learning (IRL), nonzero-sum game, vehicle formation control.**

I. Introduction

VEHICLE formation control enhances traffic efficiency, energy savings, and safety by achieving multivehicle state consensus [1], [2]. Consensus performance critically depends on reliable inter-vehicle communication. Early research, such as model predictive control (MPC) [3] and adaptive learning strategies [4], focused on control algorithm design under ideal communication conditions, laying crucial theoretical foundations. However, in practice, reliance on vulnerable wireless networks exposes formation systems to significant security threats [5], [6], [7], [8].

In practical vehicular networks, denial-of-service (DoS) attacks are prevalent and easily launched [9], [10]. By disrupting communication links, they dynamically alter the system topology, leading to formation instability and safety risks [11], [12]. Unlike traditional cyber threats, intelligent DoS attackers adapt their strategies based on defenses to maximize disruption [13]. Recent studies have addressed secure control under DoS via robust [14], [15] or H_∞ control methods [16], [17], often modeling attacks as bounded disturbances. However, these approaches largely overlook the strategic, adaptive nature of attackers, limiting their effectiveness.

Game theory effectively characterizes strategic interactions between intelligent attackers and defenders in coupled decision-making scenarios [18], [19]. By constructing game models, optimal strategies under constraints can be systematically analyzed. For instance, Li et al. [20] proposed a two-layer Stackelberg game coordinating spoofing and DoS attackers; Wu et al. [21] achieved adaptive control for heterogeneous nonlinear multiagent systems under DoS attacks via a multiplayer mixed zero-sum game. Meng et al. [22] integrated a nonzero-sum game with reinforcement learning (RL) for spacecraft attitude control. However, existing game-theoretic approaches for secure vehicle formation control face two major limitations. First, they rely on strong assumptions, making them impractical for real environments with concealed attacks [23], [24]. Second, manually predefined reward functions hinder adaptive multiobjective trade-offs, limiting system intelligence and generalization [25], [26].

In recent years, inverse RL (IRL) has become a key technology in intelligent connected vehicle control, widely applied in adaptive optimal control for complex nonlinear and multiagent systems [27], [28], [29]. Unlike conventional RL that requires manual reward function design, IRL adaptively identifies multiobjective trade-offs under adversarial conditions by learning reward signals from attack-defense demonstrations [30], [31]. The maximum entropy IRL framework [32] ensures robustness by maximizing policy entropy, while its integration with RL [33] demonstrates superior performance in complex tasks. Compared to iterative, imitation, and dual learning methods [34], [35], IRL-learned rewards offer stronger interpretability and generalizability for dynamic adversarial scenarios. However, deeply integrating IRL with nonzero-sum game models that capture real-world adversarial interactions, while constructing controllers with robust theoretical guarantees, remains an open challenge.

Based on the above analysis, this article deeply integrates the nonzero-sum game framework with IRL methods to

Received 12 February 2026; revised 13 April 2026 and 5 June 2026; accepted 1 July 2026. Date of publication 6 July 2026; date of current version 26 August 2026. This work was supported by the National Natural Science Foundation of China under Grant 62373252. *(Corresponding author: Jinliang Liu.)*

The authors are with the School of Computer Science, School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: zxp@nuist.edu.cn; guojia20202002@163.com; liujinliang@vip.163.com; wenbinhuang@nuist.edu.cn).

Digital Object Identifier 10.1109/JIOT.2026.3710354

2327-4662 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

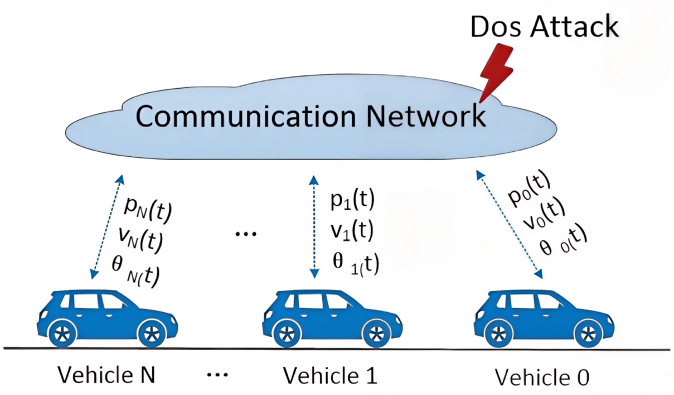

Fig. 1. Communication framework of vehicle formation under DoS attacks.

systematically address the problems of strategy modeling, reward design, and stability guarantees in vehicle formation consensus control under DoS attacks. The main contributions of this article are as follows.

1) Unlike existing studies that treat DoS attacks as random disturbances, this article models the interaction between DoS attacks and vehicle formation as a nonzero-sum game, explicitly capturing the trade-off between control performance and attack energy consumption. To overcome the reliance on manual reward tuning, this article introduces an IRL mechanism that adaptively learns multiobjective reward weights from offline attack-defense demonstration data. This enables the control system to achieve offline-adapted trade-offs among multiple objectives under varying attack conditions, without requiring manual reward engineering.
2) A neural network (NN)-based solution to the Hamilton–Jacobi (HJ) equation is developed to implement the learned reward weights in a state-feedback control law, with rigorous Lyapunov-based proof of uniformly ultimately bounded (UUB) stability for the consensus error.

II. Model Description and System Formulation

A. Graph Theory

The communication topology of the vehicle formation is abstracted as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where the node set $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ consists of N vehicles, divided into the leader node set $\mathcal{V}_l$ and the follower node set $\mathcal{V}_f$; the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ satisfies $(v_i, v_j) \in \mathcal{E}$ if vehicle i can communicate directly with vehicle j; the adjacency matrix $\mathcal{A} = [a_{ij}]_{n \times n}$ is defined such that $a_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$, otherwise $a_{ij} = 0$, with $a_{ii} = 0$ typically assumed; the degree matrix is $\mathcal{D} = \text{diag}\{d_1, d_2, \ldots, d_n\}$, where $d_i = \sum_{j=1}^{n} a_{ij}$ denotes the out-degree of node i; and the Laplacian matrix is given by $\mathcal{L} = \mathcal{D} - \mathcal{A}$.

B. System Model and Attack Model

Fig. 1 illustrates the communication status of the formation system, which consists of N vehicles under DoS attacks, where each vehicle exchanges information with its neighbors to form a closed-loop polygonal structure.

The discrete-time dynamics model of vehicle i is expressed as follows:

$$x_i(k+1) = Ax_i(k) + Bu_i(k), \quad i = 1, \ldots, N \tag{1}$$

where $x_i(k) = [p_{x_i}(k), p_{y_i}(k), \theta_i(k), v_i(k)]^T$ is the state vector of vehicle i, $p_{x_i}(k)$ and $p_{y_i}(k)$ represent the position coordinates in the global coordinate system, $\theta_i(k)$ denotes the yaw angle, and $v_i(k)$ indicates the velocity. $u_i(k)$ is the control input, and A, B are the system matrices.

Remark 1: The linear dynamics in (1) are widely adopted in formation control literature for consensus and formation problems, capturing the essential coupling between position, heading, and velocity within small deviations from the desired formation.

The global system model can be obtained as follows:

$$x(k+1) = (I_N \otimes A)x(k) + (I_N \otimes B)u(k) \tag{2}$$

where $\otimes$ is the Kronecker product and I_N is the identity matrix.

DoS attacks change the adjacency matrix by blocking the communication link. Under the attacks, the adjacency matrix elements are

$$\overline{a}_{ij}(k+1) = a_{ij}^0 \cdot (1 - \gamma_{ij}(k)) \tag{3}$$

where a_{ij}^0 is the initial adjacency matrix element, $a_{ij}^0 = 1$ and $a_{ij}^0 = 0$ indicate a connected and a disconnected link, respectively. $\gamma(k) = [\gamma_{ij}(k)]$ represents the attack strategy matrix.

The attack energy is introduced into the state as follows:

$$E(k+1) = E(k) - \sum_{i,j} \gamma_{ij}(k) \tag{4}$$

where $E(k)$ is the total attack energy at the current moment.

The Laplacian matrix after the attack is defined as follows:

$$\mathcal{L}(k+1) = \mathcal{D}(k+1) - \mathcal{A}(k+1). \tag{5}$$

At time k, the consistency error is defined based on the currently available communication topology as follows:

$$e(k) = (\mathcal{L}(k) \otimes I_m)x(k). \tag{6}$$

Similarly, at the next moment, there is

$$e(k+1) = (\mathcal{L}(k+1) \otimes I_m)x(k+1). \tag{7}$$

Substituting (2) into (7), we get

$$e(k+1) = (\mathcal{L}(k+1) \otimes A)x(k) + (\mathcal{L}(k+1) \otimes B)u(k). \tag{8}$$

Remark 2: Since the consensus error evolves solely within the disagreement subspace, and the component of the system state on the consensus manifold does not affect the error dynamics, the subsequent analysis focuses exclusively on the system behavior within the consensus error subspace.

In the consensus error subspace, the consensus error dynamics can be equivalently expressed as follows:

$$e(k+1) = \overline{A}(k)e(k) + \overline{B}(k)u(k) \tag{9}$$

where $\overline{A}(k)$ and $\overline{B}(k)$ represent the system matrices dependent on the communication topology.

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

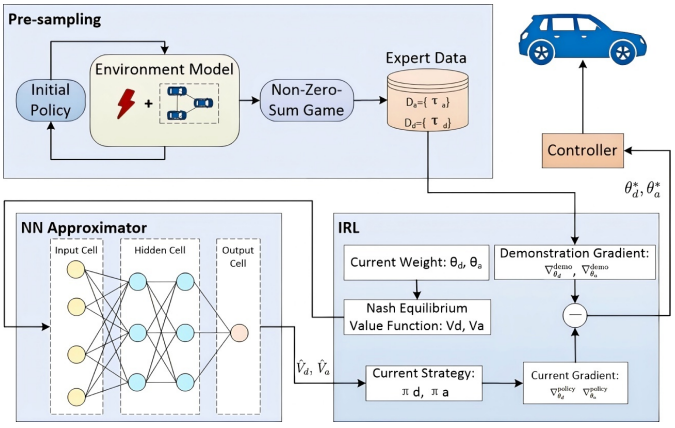

Fig. 2. Control strategy based on nonzero-sum game and IRL.

III. Main Result

This article presents an integrated nonzero-sum game and an IRL control framework. As illustrated in Fig. 2, the proposed framework achieves adaptability through a clear separation between offline learning and online feedback control. On one hand, the IRL mechanism adaptively learns the multiobjective reward weight parameters exclusively from offline demonstration data. On the other hand, the control law employs attack-intensity-aware gain scheduling that adjusts the control gains in real time according to the observed attack strength.

A. Framework of Nonzero-Sum Game

The game participants are defined as follows: 1) Defender (vehicle formation system), which aims to minimize the consensus error and control cost; and 2) Attacker (DoS attack), which aims to maximize the formation error by disrupting topology while minimizing its own energy consumption.

Lemma 1 [36]: Since both players aim to optimize their respective objectives, their Nash equilibrium satisfies

$$J_d^*\left(\pi_d^*,\pi_a^*\right) \le J_d\left(\pi_d,\pi_a^*\right), \quad \pi_d \in \mathcal{T}_d \tag{10}$$

$$J_a^*\left(\pi_d^*,\pi_a^*\right) \le J_a\left(\pi_d^*,\pi_a\right), \quad \pi_a \in \mathcal{T}_a \tag{11}$$

where $\mathcal{T}_d$ and $\mathcal{T}_a$ are the strategy sets of the defending party and the attacking party, respectively.

Lemma 2 [37]: Assume that the attacker's value function $V_a(\cdot)$ is L-Lipschitz continuous with respect to the attack variable γ_{ij}, i.e., $|V_a(\gamma_1)-V_a(\gamma_2)| \le L\|\gamma_1-\gamma_2\|$. Let $\tilde{\gamma}^* \in [0,1]^{n\times n}$ be the optimal solution to the continuous relaxation problem, and $\gamma^* = \mathrm{Proj}_{\{0,1\}}(\tilde{\gamma}^*)$ be the binary solution obtained via projection with threshold δ. Then there exists a constant C such that

$$V_a(\gamma^*) \ge V_a(\tilde{\gamma}^*) - C\cdot\Delta_{max}$$

where Δ_{max} depends on the number of relaxed variables lying in the interval $(0,1)$.

Assumption 1: To circumvent the exponential complexity of binary combinatorial optimization, this article relaxes the attack decision $\gamma_{ij} \in \{0,1\}$ to a continuous variable $\gamma_{ij} \in [0,1]$, which can be interpreted as the probability of cutting the link (i,j), thereby enabling gradient information to guide the policy search.

Theorem 1: Given the attack strategy γ, the approximately optimal defense strategy is

$$u^*(k) = -\frac{1}{2}\beta R^{-1}\overline{B}^T(k)\nabla V_d^*(e(k+1)) \tag{12}$$

where $\nabla V^*(e(k+1)) = (\partial V^*(e(k+1)))/(\partial e(k+1))$

Let $\Delta J_{ij} = \beta\nabla V_a^T(e(k+1))((\partial e(k+1))/\partial\gamma_{ij}) \le \varepsilon$. Given the defense strategy u, the heuristic attack strategy based on marginal value evaluation is

$$\gamma_{ij}^*(k) = \begin{cases} 1, & \text{if} \quad \Delta J_{ij} \ge 0 \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Proof: For subsequent calculations, write $e = e(k)$, $e^+ = e(k+1)$, and $u = u(k)$.

The defender performance indicator is defined as follows:

$$J_d(e) = \mathbb{E}\left[\sum_{t=k}^{\infty}\beta^{t-k}\left(e^TQe + u^TRu\right)\right] \tag{14}$$

where the weight matrices Q and R are positive definite, and the discount factor $\beta \in (0,1)$.

According to the principle of dynamic programming, the defensive value function satisfies

$$V_d(e) = \min_u H_d(e,u) \tag{15}$$

where $H_d(e,u) = e^TQe + u^TRu + \beta V_d(e^+)$.

Taking the partial derivative with respect to u and setting it to 0, we get

$$\frac{\partial H_d}{\partial u} = 2Ru + \beta\overline{B}(k)\nabla V_d(e^+) = 0. \tag{16}$$

Rearranging (16), we get

$$u^*(k) = -\frac{1}{2}\beta R^{-1}\overline{B}^T(k)\nabla V_d^*(e^+). \tag{17}$$

The value function of the attacker is defined as follows:

$$V_a(e) = \max_{\gamma}\sum_{\tau=k}^{\infty}\beta^{\tau-k}\left[\lambda e^TQe - \sum_{i,j}\gamma_{ij}\right] \tag{18}$$

where $\lambda > 0$ is the attacker's attention to the error.

Define the attacker's objective function as follows:

$$J_a(e) = \mathbb{E}\left[\sum_{t=k}^{\infty}\beta^{t-k}\left(\lambda e^TQe - \sum_{i,j}\gamma_{ij}\right)\right]. \tag{19}$$

Then, the Bellman equation can be expressed as follows:

$$V_a(e) = \max_{\gamma}\left[\lambda e^TQe - \sum_{i,j}\gamma_{ij} + \beta V_a(e^+)\right]. \tag{20}$$

Consider a unilateral variation in a single attack variable γ_{ij}, the resulting change in the value function can be expressed as follows:

$$\Delta V_a^{(ij)} = \lambda\Delta(e^TQe) - \Delta\gamma_{ij} + \beta\Delta V_a(e^+). \tag{21}$$

Based on Assumption 1, ignoring the immediate error change term $\lambda\Delta(e^TQe)$ and applying a first-order Taylor expansion, we have

$$\Delta V_a^{(ij)} \approx -\Delta\gamma_{ij} + \beta\nabla V_a^T(e^+)\frac{\partial e^+}{\partial\gamma_{ij}}\Delta\gamma_{ij}$$

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

$$\approx \beta \nabla V_a^T(e^+)\frac{\partial e^+}{\partial \gamma_{ij}} - 1$$
$$< \Delta J_{ij} \le \varepsilon. \tag{22}$$

Consequently, at time k, the attacker's decision on link (i, j) can be expressed as follows:

$$\gamma_{ij}^*(k) = \begin{cases} 1, & \text{if } \Delta J_{ij} \ge 0 \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

B. IRL-Based Consensus Control Strategy

The defensive reward function is defined as follows:

$$R_d(e, u; \theta_d) = \alpha_d R_1(e) + \beta_d R_2(u) \tag{24}$$

where $R_1(e) = -e^T Qe$ and $R_2(u) = -u^T Ru$, respectively, correspond to the consistency error cost and the control energy cost. $\theta_d = (\alpha_d, \beta_d)$ is the defensive reward weight parameter, satisfying $\alpha_d + \beta_d = 1$, $\alpha_d, \beta_d \ge 0$.

The attacker's reward function is defined as follows:

$$R_a(e, \gamma; \theta_a) = \xi_a R_3(e) + \zeta_a R_4(\gamma) \tag{25}$$

where $R_3(e) = e^T Qe$ and $R_4(\gamma) = -\sum \gamma_{ij}$ correspond to the attacker's reward and the energy cost, respectively. The weight parameters of the attacker's reward are denoted by $\theta_a = (\xi_a, \zeta_a)$, which satisfy $\xi_a + \zeta_a = 1$, $\xi_a, \zeta_a \ge 0$.

Let the defender's strategy $\pi_d(u|e)$ denote the probability distribution of selecting control input u, and the attacker's strategy $\pi_a(\gamma|e)$ denote the probability distribution of selecting attack decisions. Then, we get $\sum_u \pi_d(u|e) = 1$, $\sum_\gamma \pi_a(\gamma|e) = 1$.

Under the joint policy $\pi = (\pi_d, \pi_a)$, the Q-functions for the defender and the attacker are, respectively, defined as follows:

$$Q_d^\pi(e, u; \theta_d) = R_d(e, u; \theta_d) + \rho \mathbb{E}_{\pi_d}\left[V_d^\pi(e')|e, u\right] \tag{26}$$
$$Q_a^\pi(e, \gamma; \theta_a) = R_a(e, \gamma; \theta_a) + \rho \mathbb{E}_{\pi_a}\left[V_a^\pi(e')|e, \gamma\right]. \tag{27}$$

To enhance the smoothness and differentiability, the maximum entropy Nash equilibrium strategies are defined as follows:

$$\pi_d^*(u|e; \theta_d) = \frac{\exp\left(Q_d^{\pi^*}(e, u; \theta_d)\right)}{Z_d(e; \theta_d)} \tag{28}$$
$$\pi_a^*(\gamma|e; \theta_a) = \frac{\exp\left(Q_a^{\pi^*}(e, \gamma; \theta_a)\right)}{Z_a(e; \theta_a)} \tag{29}$$

where the defender's partition function is $Z_d(e; \theta_d) = \sum_u \exp(Q_a^{\pi_d^*, \pi_a^*}(e, u; \theta_d))$, and the attacker's partition function is $Z_a(e; \theta_a) = \sum_\gamma \exp(Q^{\pi_d^*, \pi_a^*}(e, \gamma; \theta_a))$.

Remark 3: In general nonzero-sum games, the maximum entropy Nash equilibrium is not guaranteed to be globally unique. Under compact strategy sets and a continuous state space, the value function is strictly concave, which ensures the existence of a local Nash equilibrium. In this article, the strategy iterative algorithm is used to solve the problem, and the stability of the strategy can be verified through multiple random initializations many times in practical applications.

According to the IRL theory, the maximum entropy optimal value function satisfies

$$V_d^{\pi^*}(e; \theta_d) = \log Z_d(e; \theta_d) \tag{30}$$
$$V_a^{\pi^*}(e; \theta_a) = \log Z_a(e; \theta_a). \tag{31}$$

Assume that the interaction process between the defender and the attacker over the discrete time interval $t = 0, 1, \ldots, T$ forms a trajectory, expressed as follows:

$$\tau = \{e_0, u_0, \gamma_0; e_1, u_1, \gamma_1; \ldots; e_T, u_T, \gamma_T\} \tag{32}$$

where e_t represents the consensus error state at time t, u_t denotes the control input applied by the defender, and γ_t indicates the attack decision.

Performing multiple independent experimental runs yields a reference trajectory set $D = \{\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(N)}\}$, based on the game interaction process. This set reflects the typical behavioral patterns of both the attacker and defender during their adversarial interactions.

The probability of a demonstration trajectory τ given the current reward weight parameters θ_d, θ_a can be expressed as follows:

$$P(\tau|\pi^*) = P(e_0) \cdot \prod_{t=0}^{T-1} \pi_d^*(u_t|e_t; \theta_d)$$
$$\pi_a^*(\gamma_t|e_t) P(e_{t+1}|e_t, u_t, \gamma_t) \tag{33}$$

where π_d^* and π_a^* denote the maximum entropy Nash equilibrium strategies corresponding to the given reward weights.

The log-likelihood objective functions for the defender's and attacker's weight parameters are, respectively, defined as follows:

$$\mathcal{L}_d(\theta_d) = \frac{1}{N} \sum_{\tau \in D} \sum_{t=0}^{T-1} \left[\log \pi_d^*(u_t|e_t; \theta_d)\right] \tag{34}$$

$$\mathcal{L}_a(\theta_a) = \frac{1}{N} \sum_{\tau \in D} \sum_{t=0}^{T-1} \left[\log \pi_a^*(\gamma_t|e_t; \theta_a)\right]. \tag{35}$$

By substituting (26), (28), and (30) into (34), the objective function can be further simplified as follows:

$$\mathcal{L}_d(\theta_d) = \frac{1}{N} \sum_{\tau \in D} \sum_{t=0}^{T-1} \left[Q_d^{\pi^*}(e_t, u_t; \theta_d) - V_d^{\pi^*}(e_t; \theta_d)\right]$$
$$= \frac{1}{N} \sum_{\tau \in D} \left[\sum_{t=0}^{T-1} R_d(e_t, u_t; \theta_d) + \rho V_d^*(e_T) - V_d^*(e_0)\right]. \tag{36}$$

When T approaches ∞, $V_d^*(e_T)$ becomes negligible, and the objective function can be approximated as follows:

$$\mathcal{L}_d(\theta_d) \approx \frac{1}{N} \sum_{\tau \in D} \left[\sum_{t=0}^{T-1} R_d(e_t, u_t; \theta_d) - V_d^{\pi^*}(e_0; \theta_d)\right]. \tag{37}$$

Remark 4: We employ the approximation $\beta V_d^*(e_T) \approx 0$ to simplify the calculation. The validity of this approximation is based on three considerations: 1) discounting; 2) the steady-state characteristics of the demonstration trajectories (the terminal error e_T is small); and 3) training stability. This simplification has been verified by subsequent simulations to have a negligible impact on the learning of the reward function weights.

Taking the partial derivative with respect to θ_d, the gradient of the objective function is obtained as follows:

$$\nabla_{\theta_d} \mathcal{L}_d = \frac{1}{N} \sum_{\tau \in D} \sum_{t=0}^{T-1} \nabla_{\theta_d} R_d(e_t, u_t; \theta_d)$$

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

$$-\frac{1}{N}\sum_{\tau\in D}\nabla_{\theta_d}V_d^{\pi^*}(e_0;\theta_d). \tag{38}$$

The value function for the defender represents the expected long-term cumulative reward starting from the initial state e_0 under the optimal policy. Its gradient can be expressed as follows:

$$\nabla_{\theta_d}V_d^{\pi^*}(e_0;\theta_d) = \mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty}\rho^t\nabla_{\theta_d}R_d(e_t,u_t;\theta_d)|e_0=e\right]. \tag{39}$$

Substituting (39) into (38), the final gradient of the defender's objective function is

$$\nabla_{\theta_d}\mathcal{L}_d = \mathbb{E}_{\tau\in D}\left[\sum_{t=0}^{T-1}\nabla_{\theta_d}R_d\right]-\mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty}\rho^t\nabla_{\theta_d}R_d\right]. \tag{40}$$

Similarly, the gradient of the attacker's objective function can be formulated as follows:

$$\nabla_{\theta_a}\mathcal{L}_a = \mathbb{E}_{\tau\in D}\left[\sum_{t=0}^{T-1}\nabla_{\theta_a}R_a\right]-\mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty}\rho^t\nabla_{\theta_a}R_a\right]. \tag{41}$$

Algorithm 1 IRL-Based Reward Parameters Learning

Initialization: defense weight $\boldsymbol{\theta}_{d,0}$ and attack weight $\boldsymbol{\theta}_{a,0}$, value functions $V_{d,0}(e)$, $V_{a,0}(e)$;
for $k=1$ to K **do**
  Compute value functions $V_{a,k}(e)$, $V_{d,k}(e)$;
  Generate maximum entropy policies $\pi_{d,k}(u|e)$, $\pi_{a,k}(\gamma|e)$;
  Compute demonstration gradients $\nabla_{\boldsymbol{\theta}_d}^{\text{demo}}$, $\nabla_{\boldsymbol{\theta}_a}^{\text{demo}}$;
  Generate sample trajectories and estimate $\nabla_{\boldsymbol{\theta}_d}^{\text{policy}}$, $\nabla_{\boldsymbol{\theta}_a}^{\text{policy}}$;
  Update reward weights $\boldsymbol{\theta}_{d,k+1}$, $\boldsymbol{\theta}_{a,k+1}$;
  if $\|\boldsymbol{\theta}_{d,k+1}-\boldsymbol{\theta}_{d,k}\|<\epsilon_1$ and $\|\boldsymbol{\theta}_{a,k+1}-\boldsymbol{\theta}_{a,k}\|<\epsilon_1$ **then**
    break
  end if
end for
Output: $\boldsymbol{\theta}_d^*$, $\boldsymbol{\theta}_a^*$;

Algorithm 1 adaptively learns the weight parameters of multiobjective performance indices from offline demonstration data, eliminating reliance on manual weight tuning.

C. Design of the NN Approximator

The optimal value functions V_d^* and V_a^*, along with their gradients, are difficult to solve analytically. Leveraging the universal approximation capability of single-hidden-layer feedforward neural networks (SLFNs), the following value function approximators are designed as:

$$\hat{V}(e)=\boldsymbol{W}^T\boldsymbol{\phi}(e)+\epsilon(e) \tag{42}$$

where $\boldsymbol{W}\in\mathbb{R}^{N_h}$ is the output weight vector; $\phi(e)=[\phi_1(e),\phi_2(e),\ldots,\phi_{N_h}(e)]^T$ is the vector of hidden layer activation functions; and $\epsilon(e)$ is the NN approximation error, satisfying $\|\epsilon(e)\|\le\epsilon_{\max}$.

Taking the gradient of $\hat{V}(e)$, we get

$$\nabla\hat{V}(e)=\nabla\boldsymbol{\phi}^T(e)\boldsymbol{W} \tag{43}$$

where $\nabla\boldsymbol{\phi}(e)$ is the Jacobian matrix of the activation function vector with respect to the error state e.

Aiming to minimize the residual of the HJ equation, the HJ residuals for the defender and the attacker are, respectively, constructed as follows:

$$\mathcal{E}_d(\boldsymbol{W}_d)=\left\|\hat{V}_d(e)-V_d^*(e)\right\| \tag{44}$$

$$\mathcal{E}_a(\boldsymbol{W}_a)=\left\|\hat{V}_a(e)-V_a^*(e)\right\|. \tag{45}$$

Construct the defender's feature matrix $\boldsymbol{\Phi}_d$ and target vector $\boldsymbol{Y}_d$ as follows:

$$\boldsymbol{\Phi}_d=\left[\boldsymbol{\phi}(e_1)-\rho\boldsymbol{\phi}(e_1'),\ldots,\boldsymbol{\phi}(e_N)-\rho\boldsymbol{\phi}(e_N')\right]^T$$
$$\boldsymbol{Y}_d=\left[e_1^TQe_1+u_1^{*T}Ru_1^*,\ldots,e_N^TQe_N+u_N^{*T}Ru_N^*\right]^T.$$

When $\boldsymbol{\Phi}_d^T\boldsymbol{\Phi}_d$ is full rank, i.e., when the persistent excitation (PE) condition is satisfied, the optimal weight for the defender is

$$\boldsymbol{W}_d^*=\left(\boldsymbol{\Phi}_d^T\boldsymbol{\Phi}_d\right)^{-1}\boldsymbol{\Phi}_d^T\boldsymbol{Y}_d. \tag{46}$$

Similarly, the attacker's feature matrix $\boldsymbol{\Phi}_a$ and target vector $\boldsymbol{Y}_a$ are constructed, respectively, as follows:

$$\boldsymbol{\Phi}_a=\left[\boldsymbol{\phi}(e_1,\gamma_1)-\rho\boldsymbol{\phi}\left(e_1',\gamma_1'\right),\ldots,\boldsymbol{\phi}(e_N,\gamma_N)-\rho\boldsymbol{\phi}\left(e_N',\gamma_N'\right)\right]^T$$
$$\boldsymbol{Y}_a=\left[\lambda e_1^TQe_1-\sum\gamma_{1ij},\ldots,e_n^TQe_n-\sum\gamma_{Nij}\right]^T.$$

The optimal weight for the attacker is

$$\boldsymbol{W}_a^*=\left(\boldsymbol{\Phi}_a^T\boldsymbol{\Phi}_a\right)^{-1}\boldsymbol{\Phi}_a^T\boldsymbol{Y}_a. \tag{47}$$

Algorithm 2 NN for Approximating Value Functions

Initialization: weight vectors $\boldsymbol{W}_{d,0}$, $\boldsymbol{W}_{a,0}$; initial policies $u_0=0$, $\gamma_0=0$; convergence thresholds ϵ_1, ϵ_2;
for $k=1$ to K **do**
  Compute gradients $\nabla\hat{V}_{d,k}$, $\nabla\hat{V}_{a,k}$;
  Update policies u_{k+1}^*, γ_{k+1}^*;
  Generate trajectory data using new policies;
  Solve $\boldsymbol{W}_{d,k+1}$, $\boldsymbol{W}_{a,k+1}$ using batch least squares;
  if $\|\boldsymbol{W}_{d,k+1}-\boldsymbol{W}_{d,k}\|<\epsilon_2$, $\|\boldsymbol{W}_{a,k+1}-\boldsymbol{W}_{a,k}\|<\epsilon_2$ **then**
    break
  end if
end for
Output: $\boldsymbol{W}_d^*$, $\boldsymbol{W}_a^*$;

Algorithm 2 approximates the optimal value functions in the differential game, alternately updating the strategy and network weights until convergence to output the optimal weights.

Remark 5: The PE condition required for $\Phi^T\Phi$ to be full rank can be ensured in practice by introducing exploration noise into the control input during data collection, or by using

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

sufficiently rich excitation trajectories. This guarantees sufficient state-space coverage for reliable parameter estimation.

D. Stability Analysis

Theorem 2: If the attack energy γ^*, the defender strategy u^*, and the NN approximation error $\epsilon(e)$ are bounded, the consensus error dynamics system in (9) is globally UUB stable in the closed loop. That is, there exists a constant $M > 0$ such that

$$\lim_{k\to\infty} \sup \|e(k)\| \leq \sqrt{\frac{\epsilon_2}{\rho k_1}} = M. \tag{48}$$

Proof: Select the defender's approximately optimal value function as the Lyapunov function

$$V(e(k)) = V_d'(e(k)) = \boldsymbol{W}_d^T \boldsymbol{\phi}(e(k)) \tag{49}$$

where $V_d'(e)$ is the NN's approximation of the defender's optimal value function $V_d^*(e)$, satisfying

$$V_d^*(e) = V_d'(e) + \epsilon(e), \quad |\epsilon(e) \leq \epsilon_2|. \tag{50}$$

Since the instantaneous cost function is positive definite and quadratic, and the value function represents the cumulative discounted cost, there exist positive constants c_1 and c_2 such that

$$c_1\|e\|^2 \leq V_d^*(e) \leq c_2\|e\|^2. \tag{51}$$

Under the condition of bounded approximation error, (51) still holds for $V_d'(e)$.

Combining (15), (17), and (50), we have

$$\begin{aligned} V_d'(e(k)) = {} & e(k)^T Q e(k) + u^*(k)^T R u^*(k) \\ & + \rho V_d'(e(k+1)) + \epsilon(e). \end{aligned} \tag{52}$$

Rearranging (52), we obtain

$$\rho V_d'(e(k+1)) - V_d'(e(k)) \leq -\lambda_{min}(Q)\|e(k)\|^2 + \epsilon_2. \tag{53}$$

Define $\Delta V = V_d'(e(k+1)) - V_d'(e(k))$, we have

$$\begin{aligned} \rho\Delta V &= \rho(V_d'(e(k+1)) - V_d'(e(k))) + (1-\rho)V_d'(e(k)) \\ &\leq -\lambda_{min}(Q)\|e(k)\|^2 + \epsilon_2 + (1-\rho)c_2\|e(k)\|^2. \end{aligned} \tag{54}$$

Then, we get

$$\Delta V \leq -\frac{\lambda_{min}(Q) - (1-\rho)c_2}{\rho}\|e(k)\|^2 + \frac{\epsilon_2}{\rho}. \tag{55}$$

Assume $\lambda_{min}(Q) > (1-\rho)c_2$, and define a positive constant

$$k_1 = \frac{\lambda_{min}(Q) - (1-\rho)c_2}{\rho} > 0. \tag{56}$$

Then, the Lyapunov difference becomes

$$\Delta V \leq -k_1\|e(k)\|^2 + \frac{\epsilon_2}{\rho}. \tag{57}$$

When k tends to infinity, we have

$$\lim_{k\to\infty} sup\|e(k)\| \leq \sqrt{\frac{\epsilon_2}{\rho k_1}}. \tag{58}$$

Remark 6: The derived upper bound indicates that the consensus error remains within a compact set determined by the attack intensity and approximation error. In practical vehicle formation scenarios, this bound can be used to ensure that inter-vehicle distances remain within safety margins by appropriately tuning the weighting matrices Q and R.

Theorem 3: Under the maximum entropy policies (28) and (29), the log-likelihood $\mathcal{L}(\theta)$ satisfies.

1) *Concavity:* $\nabla_\theta^2\mathcal{L}(\theta) = -\mathbb{E}[\mathrm{Var}_{\pi^*}(\nabla_\theta Q^\pi)] \leq 0$.
2) *Unbiasedness:* $\mathbb{E}[\hat{\nabla}_\theta\mathcal{L}] = \nabla_\theta\mathcal{L}$.
3) *Decoupling:* $\nabla_{\theta_d}\mathcal{L}_d$ depends only on θ_d, $\nabla_{\theta_a}\mathcal{L}_a$ only on θ_a.

Hence, $\theta_{k+1} = \theta_k + \eta_k\hat{\nabla}_\theta\mathcal{L}(\theta_k)$ with $\sum\eta_k = \infty$, $\sum\eta_k^2 < \infty$ converges a.s. to a local maximum.

Proof: From (24), the reward function is linear in θ_d

$$R_d(e,u;\theta_d) = \theta_d^T r_d(e,u) \tag{59}$$

where $r_d(e,u) = [R_1(e), R_2(u)]^T$. Consequently, the Q-function satisfies $Q_d^\pi(e,u;\theta_d) = \theta_d^T\psi(e,u)$ with $\psi(e,u) = r_d(e,u) + \rho\mathbb{E}[V_d^\pi(e')|e,u]$.

The log-likelihood objective is

$$\mathcal{L}_d(\theta_d) = \mathbb{E}_{\tau\sim D}\left[\log\pi_d^*(u|e;\theta_d)\right] \tag{60}$$

where $\pi_d^*(u|e;\theta_d) = \exp(Q_d^\pi(e,u;\theta_d) - \log Z_d(e;\theta_d))$ and $Z_d(e;\theta_d) = \sum_u \exp(Q_d^\pi(e,u;\theta_d))$.

The Hessian of $\mathcal{L}_d$ is

$$\nabla_{\theta_d}^2\mathcal{L}_d = -\mathbb{E}_{(e,u)\sim D}\left[\mathrm{Var}_{\pi_d^*(\cdot|e;\theta_d)}(\psi(e,u))\right] \leq 0 \tag{61}$$

where $\mathrm{Var}_{\pi^*}(\cdot)$ denotes the covariance matrix under π^*, which is positive semidefinite. Thus, $\mathcal{L}_d$ is concave. The same argument applies to $\mathcal{L}_a(\theta_a)$ by symmetry.

The gradient estimate in (40) consists of two independent parts

$$\begin{aligned} \mathbb{E}[\hat{\nabla}_{\theta_d}\mathcal{L}_d] &= \mathbb{E}_{\tau\sim D}\left[\sum\nabla_{\theta_d}R_d\right] - \mathbb{E}_{\pi^*}\left[\sum\rho^t\nabla_{\theta_d}R_d\right] \\ &= \nabla_{\theta_d}\mathcal{L}_d. \end{aligned} \tag{62}$$

Hence, the gradient estimate is unbiased.

In the nonzero-sum game formulation, the defender and attacker have independent reward structures defined in (24) and (25), respectively. Specifically, $\theta_d = (\alpha_d, \beta_d)$ appears only in R_d, and $\theta_a = (\xi_a, \zeta_a)$ appears only in R_a. Consequently, $\nabla_{\theta_d}\mathcal{L}_d$ depends solely on θ_d, and $\nabla_{\theta_a}\mathcal{L}_a$ depends solely on θ_a. The gradient updates for the two players are completely decoupled.

In conclusion, the stochastic gradient ascent update satisfies the standard Robbins-Monro conditions, ensuring almost sure convergence to a local maximum.

IV. Simulation Example

A. Parameter Setting

The simulation adopts a three-vehicle triangular formation, a representative case that can be readily extended to the same ring-type communication topology and control architecture when more vehicles are added.

The simulation experiments are implemented in MATLAB R2022a. The system consists of a triangular formation with one leader and two followers. The vehicle dynamics follow

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

the discrete-time linear model in (1), with its system matrices configured as follows:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0.1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

In the formation system, the leader vehicle is located at the origin $(0, 0)$, the left-wing vehicle is at $(-15, 12)$, and the right-wing vehicle is at $(-15, -12)$. The communication topology adopts a directed chain structure, with its adjacency matrix defined as $[0, 1, 0; 1, 0, 1; 1, 1, 0]$.

In the value function and reward function, the weighting matrices are set as $Q = diag(1, 1, 0.1, 0.5)$, $R = diag(0.1, 0.05)$. The discount factor $\rho = 0.95$ and attacker's error concern factor $\lambda = 1.5$ are selected from [37]. Unlike the zero-sum IRL setting in [37], our nonzero-sum formulation allows both players to have independent reward structures, which is essential for modeling DoS attacks where the attacker cares about both disruption and energy consumption. Furthermore, while [37] assumes known rewards, we learn the reward weights adaptively from attack-defense demonstration trajectories using maximum-entropy IRL.

To satisfy the PE condition, zero-mean Gaussian noise with variance $\sigma^2 = 0.05$ is added to the control input during the data collection phase.

To comprehensively evaluate performance in a network environment subject to DoS attacks, this article considers a DoS attack model with energy caps $E_{\max}$ set to 30, 60, and 100, respectively. The proposed method (G-IRL) is compared against the method based on MPC, game + RL (G-RL), H_∞ control, and robust control (RC).

The demonstration trajectory set for both the attacker and defender contains 100 trajectories, where $p_x, p_y \in [-2, 2]$ m, $\theta \in [-0.5, 0.5]$ rad, and $v \in [0, 2]$ m/s. The specific numerical characteristics of the reference trajectory set are as follows: the mean initial formation error is 0.85 m with a standard deviation of 0.42 m; the amplitude range of the defender's control input is $[-0.35, 0.35]$. Through quality filtering, only high-quality trajectories with an average consensus error below 1.0 m are retained as the reference trajectories.

B. Result Discussion

Fig. 3 illustrates the dynamic coupling between DoS attack intensity and defense control input: the defender's input (red dashed) adjusts synchronously with the number of attacked links (blue solid), demonstrating that the approximate Nash equilibrium strategy adaptively modulates control effort based on attack behavior, validating the game-theoretic modeling.

Fig. 4 shows that under weak to strong DoS attacks, the proposed method (red solid) consistently outperforms other methods in both convergence speed and steady-state error, demonstrating its enhanced robustness. This result validates that our method effectively enhances formation consensus performance and robustness against DoS attacks.

Fig. 5 shows that G-IRL maintains the lowest and most stable defense energy consumption across all attack intensities,

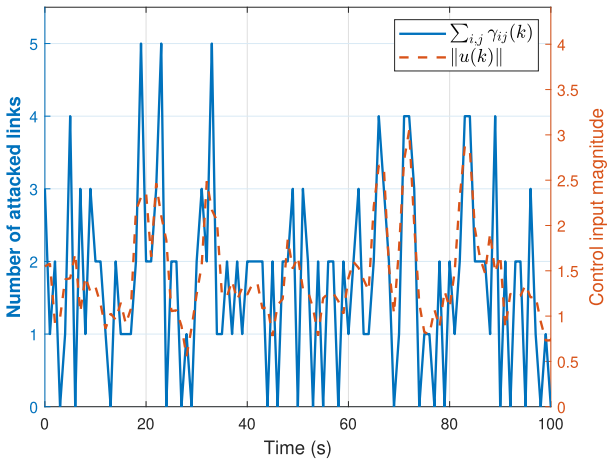

Fig. 3. Evolution of attack and defense strategies.

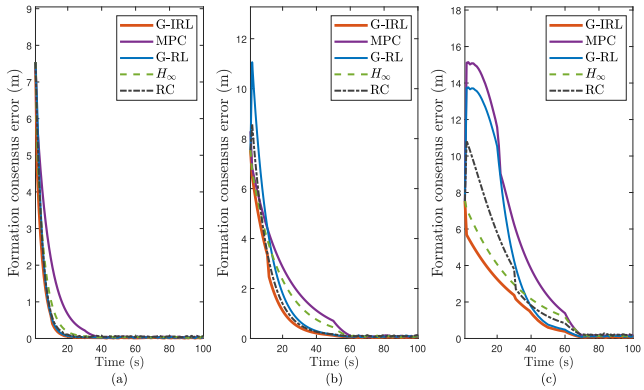

Fig. 4. Comparison of global consensus error convergence of different methods under different DoS attack intensities. (a) $E_{max} = 30$. (b) $E_{max} = 60$. (c) $E_{max} = 100$.

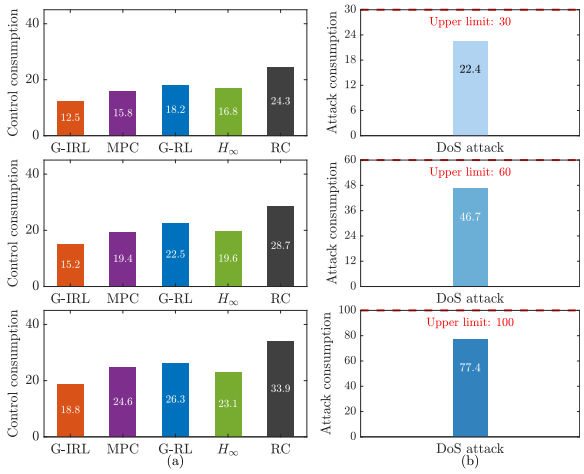

Fig. 5. Analysis of attack and defense energy consumption under different DoS attack intensities. (a) Defense energy consumption. (b) Attack energy consumption.

confirming that it effectively balances defense performance and energy consumption while mitigating DoS attack impacts.

Fig. 6 demonstrates that the leader and followers maintain a stable triangular formation along an S-shaped path with smooth trajectories and no oscillations. This demonstrates that the proposed method retains good tracking performance

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

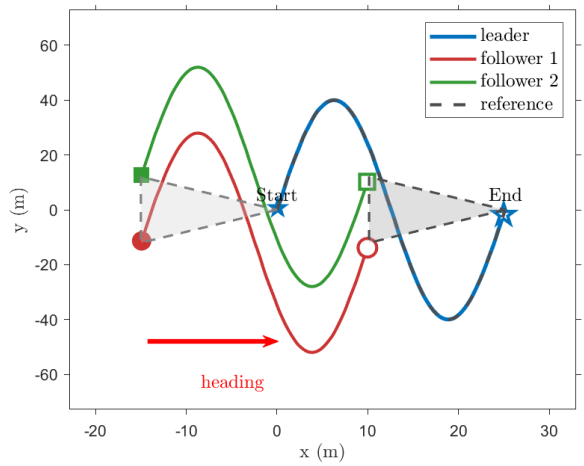

Fig. 6. Trajectory tracking performance of vehicle formation.

TABLE I
SENSITIVITY ANALYSIS OF REWARD WEIGHTS ABOUT T AND ρ

T	ρ	$\alpha_d(mean \pm std)$	$\xi_a(mean \pm std)$
50	0.90	0.6492 ± 0.0038	0.5675 ± 0.0076
50	0.95	0.6622 ± 0.0049	0.5860 ± 0.0061
50	0.99	0.6673 ± 0.0043	0.6021 ± 0.0099
100	0.90	0.6443 ± 0.0086	0.5695 ± 0.0076
100	0.95	0.6676 ± 0.0135	0.5920 ± 0.0130
100	0.99	0.6673 ± 0.0085	0.6021 ± 0.0116
200	0.90	0.6453 ± 0.0150	0.5674 ± 0.0053
200	0.95	0.6757 ± 0.0105	0.5984 ± 0.0089
200	0.99	0.6824 ± 0.0103	0.5969 ± 0.0117

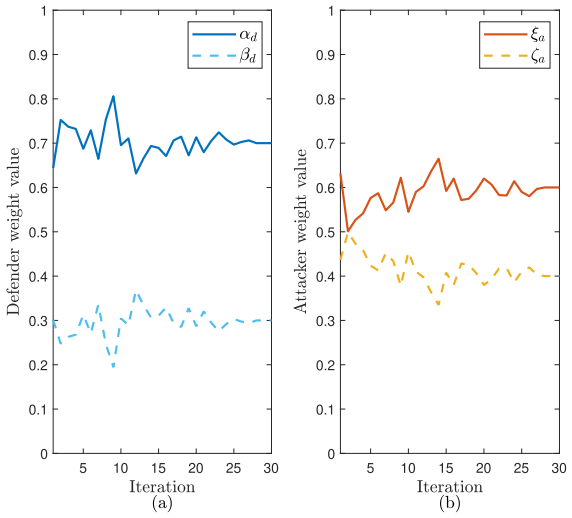

Fig. 7. Iterative convergence characteristics of reward weights of defender and attacker in IRL. (a) Defender weight value. (b) Attacker weight value.

and formation-keeping capability even under complex path conditions.

As shown in Table I, the learned reward weights exhibit only minor variations when T varies from 50 to 200 and ρ from 0.90 to 0.99, confirming that the terminal approximation in (37) introduces no significant bias.

Fig. 7 shows that a higher weight on error prioritizes formation precision, while a higher weight on control cost

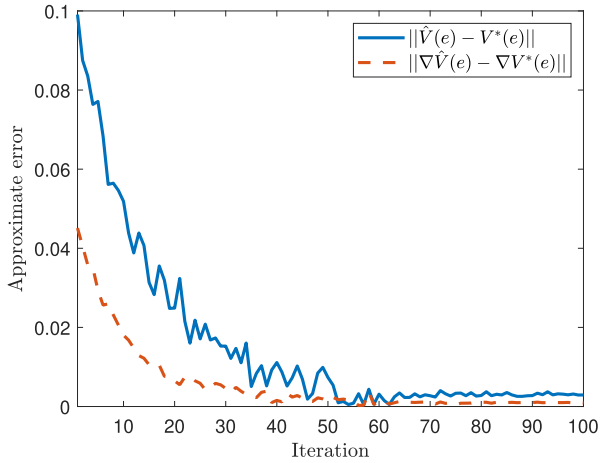

Fig. 8. Convergence characteristics of NN approximation error for the value function.

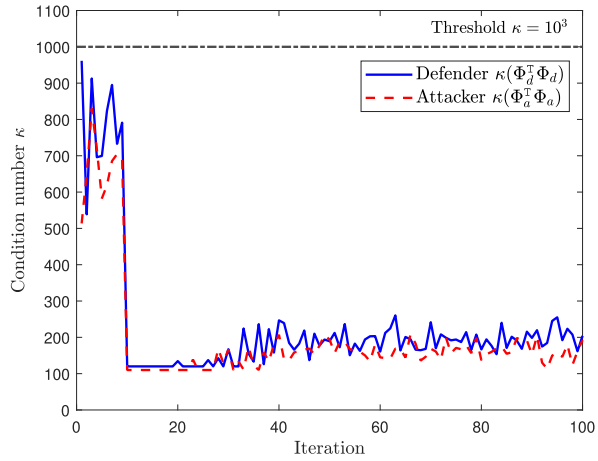

Fig. 9. Convergence of the condition number $\kappa(\cdot)$ over training iterations.

favors energy saving. This adjustment enables the system to balance safety and efficiency under varying attack conditions.

Fig. 8 demonstrates that the NN approximation errors for the value function and its gradient decrease rapidly and stabilize at low levels, ensuring a reliable HJ equation solution.

In Fig. 9, the condition numbers of both the defender and attacker feature matrices converge to below 10^3 (approximately 190 and 170, respectively). In numerical linear algebra, this indicates that the feature matrices are well-conditioned and the least-squares solutions are numerically stable [38].

Fig. 10 shows that under actuator saturation (E_{max} = 60), the proposed G-IRL method (red solid) achieves a steady-state error of 0.36 m after 60 s, significantly lower than other saturated controllers, while the unsaturated G-IRL (gray dashed) attains 0.25 m as a benchmark, demonstrating that IRL-learned weights enable superior long-term convergence despite initial saturation-induced degradation.

V. CONCLUSION

This article proposes a G-IRL control framework that demonstrates superior convergence and robustness against DoS attacks by modeling attacker-defender interactions and adaptively learning multiobjective weights from offline

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

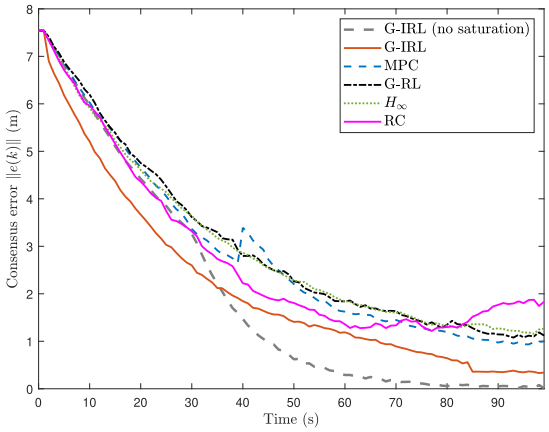

Fig. 10. Consensus error under actuator saturation.

demonstrations. However, limitations of the current study include the adoption of a linearized system model and static hyperparameters. Among various nonlinear effects, actuator saturation is the most critical for vehicle formation control, which would lead to moderate performance degradation. Therefore, future work will aim to accommodate nonlinear vehicle dynamics by incorporating NN-based dynamics models to improve the framework's resilience and autonomy in dynamic scenarios.

REFERENCES

[1] C. Wu, Z. Cai, Y. He, and X. Lu, "A review of vehicle group intelligence in a connected environment," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1865–1889, Jan. 2024.

[2] V. Lesch, M. Breitbach, M. Segata, C. Becker, S. Kounev, and C. Krupitzer, "An overview on approaches for coordination of platoons," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10049–10065, Aug. 2022.

[3] Y. Liu, Z. Li, Q. Li, and X. Xie, "Secure consensus control for connected vehicle systems with resilient predictors against denial-of-service attacks," *IEEE Access*, vol. 12, pp. 41908–41917, 2024.

[4] J. Huang, W. Wang, and X. Su, "Adaptive iterative learning control of multiple autonomous vehicles with a time-varying reference under actuator faults," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5512–5525, Dec. 2021.

[5] S. Hussain, S. Tahir, A. Masood, and H. Tahir, "Blockchain-enabled secure communication framework for enhancing trust and access control in the Internet of Vehicles (IoV)," *IEEE Access*, vol. 12, pp. 110992–111006, 2024.

[6] C. Chu, Y. Cao, B. Niu, N. Zhao, and L. Zhang, "Human-in-the-loop leader-following consensus control for nonlinear MASs subject to deception attacks via dynamic self-triggered mechanism," *IEEE Syst. J.*, vol. 19, no. 4, pp. 1088–1098, Dec. 2025.

[7] F. Yang, J. Liu, Q. Ma, and X. Xie, "Dynamic event-triggered predefined-time distributed optimal consensus for multiple Euler–Lagrangian systems with unavailable parameters," *IEEE Syst. J.*, vol. 19, no. 2, pp. 518–528, Jun. 2025.

[8] J. Liu, E. Ma, L. Zha, and E. Tian, "Distributed energy management for microgrids: When privacy preservation meets multiple mechanisms," *IEEE Trans. Consum. Electron.*, vol. 72, no. 1, pp. 1514–1523, Feb. 2026.

[9] H. Liu and L. Hao, "An improved data-driven iterative learning secure control for intelligent marine vehicles with DoS attacks," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 2160–2170, Jan. 2023.

[10] J. Liu, Z. Liu, Y. Li, E. Tian, and C. Peng, "Privacy-protected formation control for unmanned surface vehicles: A neural predictor-based non-cooperative game framework," *IEEE Trans. Veh. Technol.*, early access, Jan. 12, 2026, doi: 10.1109/TVT.2026.3651837.

[11] C. Liu, Z. Xia, and R. J. Patton, "Distributed fault-tolerant consensus control of vehicle platoon systems with DoS attacks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 10, pp. 14438–14449, Oct. 2024.

[12] J. Wang, J. H. Park, X. Li, L. Jin, and X. Luo, "Resilient DMPC-based string stable platoon control under DoS attacks," *IEEE Internet Things J.*, vol. 12, no. 20, pp. 41690–41701, Oct. 2025.

[13] G. Zhao, Y. Ning, and H. Cui, "Consensus of multi-agent systems with DoS attacks and actuator faults: A double-layer event-triggered control approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 23288–23300, 2025.

[14] Q. Han, J. Ma, Z. Zuo, X. Wang, B. Yang, and X. Guan, "Resilient event-triggered control of vehicle platoon under DoS attacks and parameter uncertainty," *IEEE Trans. Intell. Vehicles*, vol. 10, no. 1, pp. 169–178, Jan. 2025.

[15] X. Song, X. Liu, H. Shan, and R. Lu, "Time-delay robust safety control of connected vehicles under stochastic DoS attack," *IEEE Internet Things J.*, vol. 12, no. 15, pp. 29485–29494, Aug. 2025.

[16] S. M. Elkhider, S. El-Ferik, and A. A. Saif, "Containment control of multiagent systems subject to denial of service attacks," *IEEE Access*, vol. 10, pp. 48102–48111, 2022.

[17] Z. Gu, X. Sun, H.-K. Lam, D. Yue, and X. Xie, "Event-based secure control of T–S fuzzy-based 5-DOF active semivehicle suspension systems subject to DoS attacks," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 2032–2043, Jun. 2022.

[18] F. Tan and K. Qi, "Distributed non-cooperative games and distributed learning in linear and nonlinear systems: An overview," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 71, no. 8, pp. 3843–3856, Aug. 2024.

[19] H. B. Jond and J. Platoć, "Differential game-based optimal control of autonomous vehicle convoy," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 2903–2919, Mar. 2023.

[20] L. Li, Y. Li, J. Lian, and M. Li, "Active security control for switched systems under deception and DoS attacks based on two-tier Stackelberg game," *IEEE Trans. Cybern.*, vol. 55, no. 7, pp. 3251–3261, Jul. 2025.

[21] Y. Wu, M. Chen, H. Li, and M. Chadli, "Mixed-zero-sum-game-based memory event-triggered cooperative control of heterogeneous MASs against DoS attacks," *IEEE Trans. Cybern.*, vol. 54, no. 10, pp. 5733–5745, Oct. 2024.

[22] Y. Meng, X. Yan, J. Huang, and D. Zhu, "Optimal and robust spacecraft attitude control in complex constrained environments via non-zero-sum game theory," in *Proc. 37th Chin. Control Decis. Conf. (CCDC)*, May 2025, pp. 5438–5443.

[23] A. Signori et al., "A geometry-based game theoretical model of blind and reactive underwater jamming," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3737–3751, Jun. 2022.

[24] J. Tang, S. Shao, J. Song, and A. Gupta, "Nash equilibrium control policy against bus-off attacks in CAN networks," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 980–990, 2023.

[25] Y. Jiang and Z. Li, "Fully distributed target encircling control of autonomous surface vehicles based on noncooperative games," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 4, pp. 4769–4779, Apr. 2024.

[26] R. Tian, N. Li, I. Kolmanovsky, Y. Yildiz, and A. R. Girard, "Game-theoretic modeling of traffic in unsignalized intersection network for autonomous vehicle control verification and validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2211–2226, Mar. 2022.

[27] P. Li, L. Fu, Z. Song, and Z. Wang, "Predictor-based event-triggered optimized control for uncertain nonlinear systems with time delays: A reinforcement learning approach," *IEEE Trans. Fuzzy Syst.*, vol. 33, no. 9, pp. 3360–3374, Sep. 2025.

[28] M. Ran and L. Xie, "Adaptive observation-based efficient reinforcement learning for uncertain systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5492–5503, Oct. 2022.

[29] J. Liu, Z. Zhang, E. Tian, C. Peng, J. Cao, and T. Huang, "Reinforcement learning-based formation control for uncrewed surface vehicles under aperiodic DoS attacks: A Stackelberg–Nash game approach," *IEEE Trans. Cybern.*, early access, Mar. 27, 2026, doi: 10.1109/TCYB.2026.3674952.

[30] W. Xue, B. Lian, Y. Kartal, J. Fan, T. Chai, and F. L. Lewis, "Model-free inverse H-infinity control for imitation learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 5661–5672, 2025.

[31] J. Nan, R. Zhang, G. Yin, W. Zhuang, Y. Zhang, and W. Deng, "Safe and interpretable human-like planning with transformer-based deep inverse reinforcement learning for autonomous driving," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 12134–12146, 2025.

[32] F. M. Golmisheh and S. Shamaghdari, "Optimal robust formation of multi-agent systems as adversarial graphical apprentice games with inverse reinforcement learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 4867–4880, 2024.

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.

[33] W. Li, F. Qiu, L. Li, Y. Zhang, and K. Wang, "Simulation of vehicle interaction behavior in merging scenarios: A deep maximum entropy-inverse reinforcement learning method combined with game theory," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1079–1093, Jan. 2023.
[34] F. Yang, Z. Gong, Q. Wei, and Y. Lei, "Secure containment control for multi-UAV systems by fixed-time convergent reinforcement learning," *IEEE Trans. Cybern.*, vol. 55, no. 4, pp. 1981–1994, Apr. 2025.
[35] J. An, L. Cao, Y. Wang, A. Khan Jadoon, and S. Wang, "Adaptive fault-tolerant optimized platoon cloud tracking control for heterogeneous vehicles via dual learning mechanism," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 4382–4393, 2025.
[36] W. Li, Q. Li, S. E. Li, R. Li, Y. Ren, and W. Wang, "Indirect shared control through non-zero sum differential game for cooperative automated driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15980–15992, Sep. 2022.
[37] X. Lin, P. A. Beling, and R. Cogill, "Multiagent inverse reinforcement learning for two-person zero-sum games," *IEEE Trans. Games*, vol. 10, no. 1, pp. 56–68, Mar. 2018.
[38] X. Dingyü, *Linear Algebra and Matrix Computations With MATLAB*. Berlin, Germany: De Gruyter, 2020.

Xiaoping Zhao received the Ph.D. degree in measurement technology and instruments from the Institute of Vibration Engineering, School of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in March 2009.

Since April 2009, she has been with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, where she is currently an Associate Professor. From February 2015 to August 2015, she was a Visiting Scholar with New Jersey Institute of Technology, Newark, NJ, USA. Her current research interests include connected vehicles, autonomous surface vessels, autonomous surface craft, and deep learning.

Jia Guo received the B.S. degree in information and computing science from Changshu Institute of Technology, Suzhou, China, in 2020. She is currently pursuing the master's degree with Nanjing University of Information Science and Technology, Nanjing, China.

Her research interests include connected vehicle control systems, security control, and privacy protection.

Jinliang Liu (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the School of Information Science and Technology, Donghua University, Shanghai, China, in 2011.

From December 2013 to June 2016, he was a Post-Doctoral Research Associate with the School of Automation, Southeast University, Nanjing, China. From October 2016 to October 2017, he was a Visiting Researcher/Scholar with the Department of Mechanical Engineering, The University of Hong Kong, Hong Kong. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical Engineering, Yeungnam University, Gyeongsan, South Korea. From 2011 to 2023, he was an Associate Professor and then a Professor with Nanjing University of Finance and Economics, Nanjing. In 2023, he joined Nanjing University of Information Science and Technology, Nanjing, where he is currently a Professor with the School of Computer Science. His research interests include cyber-physical systems, autonomous systems, privacy preservation, and game theory.

Wenbin Huang received the Ph.D. degree from Hunan University, Changsha, China, in 2023.

He currently serves as an Associate Professor at the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China. He has published multiple papers in conferences and journals, such as IEEE INFOCOM, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE INTERNET OF THINGS JOURNAL (IOT-J), and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. His primary research focus is on IoT security and privacy.

Authorized licensed use limited to: Nanjing University of Information Science and Technology. Downloaded on August 30,2026 at 14:24:37 UTC from IEEE Xplore. Restrictions apply.