mathematics

MDPI

Article

Cooperative Guidance Strategy for Active Spacecraft Protection from a Homing Interceptor via Deep Reinforcement Learning

Weilin Ni [1], Jiaqi Liu [2], Zhi Li [1], Peng Liu [2] and Haizhao Liang [1,*]

[1] School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzheng 518107, China; niwlin@mail2.sysu.edu.cn (W.N.); lizh336@mail2.sysu.edu.cn (Z.L.)
[2] National Key Laboratory of Science and Technology on Test Physics and Numerical Mathematics, Beijing 100076, China; liujiaqi_business@163.com (J.L.); lppl2008@163.com (P.L.)
* Correspondence: lianghch5@mail.sysu.edu.cn; Tel.: +86-133-211-967-99

Abstract: The cooperative active defense guidance problem for a spacecraft with active defense is investigated in this paper. An engagement between a spacecraft, an active defense vehicle, and an interceptor is considered, where the target spacecraft with active defense will attempt to evade the interceptor. Prior knowledge uncertainty and observation noise are taken into account simultaneously, which are vital for traditional guidance strategies such as the differential-game-based guidance method. In this set, we propose an intelligent cooperative active defense (ICAAI) guidance strategy based on deep reinforcement learning. ICAAI effectively coordinates defender and target maneuvers to achieve successful evasion with less prior knowledge and observational noise. Furthermore, we introduce an efficient and stable convergence (ESC) training approach employing reward shaping and curriculum learning to tackle the sparse reward problem in ICAAI training. Numerical experiments are included to demonstrate ICAAI's real-time performance, convergence, adaptiveness, and robustness through the learning process and Monte Carlo simulations. The learning process showcases improved convergence efficiency with ESC, while simulation results illustrate ICAAI's enhanced robustness and adaptiveness compared to optimal guidance laws.

Keywords: cooperative guidance; reinforcement learning; active protection; guidance law

MSC: 93-08

Citation: Ni, W.; Liu, J.; Li, Z.; Liu, P.; Liang, H. Cooperative Guidance Strategy for Active Spacecraft Protection from a Homing Interceptor via Deep Reinforcement Learning. *Mathematics* **2023**, *11*, 4211. https://doi.org/10.3390/math11194211

Academic Editor: Jiangping Hu

Received: 29 August 2023
Revised: 27 September 2023
Accepted: 7 October 2023
Published: 9 October 2023

1. Introduction

Spacecraft such as satellites, space stations, and space shuttles play an important role in both civil and military activities. They are also at risk of being intercepted in the exo-atmosphere. The pursuit-evasion game between the spacecraft and the interceptor will be critical in the competition for space resources and has been widely studied in recent years. The trajectory of spacecraft can be accurately predicted [1] since the dynamics of the spacecraft is generally described in terms of a two-body problem. With the development of accurate sensors, guidance technology, small-sized propulsion systems, and fast servo-mechanism techniques, the Kinetic Kill Vehicle (KKV), which can be used for direct-hit killing, has superior maneuverability compared to the other spacecraft. In other words, it is not practical for targeted spacecraft involved in the pursuit-evasion game to rely solely on orbital maneuvering.

Among the many available countermeasures, launching an Active Defense Vehicle (ADV) as a defender to intercept the incoming threat has proven to be an effective approach to compensate for the inferior target maneuverability [2–4]. In an initial study [2], Boyell proposed the active defense strategy of launching a defensive missile to protect the target from a homing missile. Boyell proposed an approximate normalized curve of game results under the condition of constant or static target velocity based on the relative motion relationship among the three participants. The dynamic three-body framework was

introduced by Rusnak in Ref. [4], inspired by the narrative of a "lady-bodyguard-bandit" situation. This framework was later transformed into a "target-interceptor-defender" (TID) three-body spacecraft active defense game scenario as described in Ref [3]. In the TID scenario, the defender aims to reduce the distance from the interceptor, while the interceptor endeavors to increase the distance from the defender and successfully intercept the target. In Refs. [3,4], Rusnak proposed a game guidance method under the TID scenario based on Multiple Objective Optimization and differential games theories. It was proven that the proposed active defense method significantly reduces the miss distance and the required acceleration level between interceptor and defender.

The efficacy of the active defense method has garnered increased attention to the collaborative strategy between the target and defender in the TID scenario. Traditional methods for solving optimal strategies in this context include Optimal Control [5–7] and differential games theories [8–10]. In Ref. [7], Weiss employed the Optimal Control theory to independently design the guidance for both the target and defender. This approach considered the influence of target maneuvers on the interceptor's effectiveness as a defender. Furthermore, in Ref. [6], collaborative game strategies for the target and defender were proposed, emphasizing their combined efforts in the TID scenario. Aiming at the multi-member TID scenario in which a single target carries two defenders against two interceptors, Ref. [5] designed a multi-member cooperative game guidance strategy and considered the fuel consumption of target and defender. However, Optimal-Control-based strategies rely on perfect information, demanding accurate maneuvering details of the interceptor. In contrast, Differential Game approaches require prior knowledge instead of accurate target acceleration information, enhancing algorithm robustness [11]. In Ref. [8], optimal cooperative pursuit and evasion strategies were proposed using Pontryagin's minimum principle. A similar scenario was studied in Ref. [9] for both continuous and discrete domains using the linear–quadratic differential game method. It is worth noting that the differential game control strategies proposed in Ref. [9] solve the fuel cost and saturation problem. However, they introduce computational problems and make the selection of weight parameters more difficult. A switching surface [10], designed with zero-effort miss distance, was introduced to divide the multi-agent engagement into two one-on-one differential games, thereby achieving a balance between performance and usability. Nonetheless, using the differential game method to solve the multi-agent pursuit-evasion game problem still faces shortcomings [11–13]. First, it is difficult to establish a scene model of a multi-member, multi-role game due to the extremely large increase in the dimension of the state quantity; second, it has high requirements for the accuracy of the prior knowledge, and the success rate of the game is low if the prior knowledge of the players in the game cannot be obtained accurately; third, the differential game algorithm is complicated, involving a high-dimensional matrix operation, power function operation, integral calculation, etc., which places a high demand on the computational resources of the spacecraft. More on this topic can be found in [14–20].

With the advancement of machine learning technology, Deep Reinforcement Learning (DRL) has emerged as a promising approach for addressing active defense guidance problems. In DRL, an agent interacts with the environment and receives feedback in the form of rewards, enabling it to improve its performance and achieve specific tasks. This mechanism has led to successful applications of DRL in various decision-making domains, including robot control, MOBA games, autonomous driving, and navigation [21–25]. In Ref. [26], the DRL was utilized to learn an adaptive homing phase control law, accounting for sensor and actuator noise and delays. Another work [27] proposed an adaptive guidance system to address the landing problem using Reinforcement Meta-Learning, adapting agent training from one environment to another with limited steps, showcasing robust policy optimization in the presence of parameter uncertainties. In the context of the TID scenario, Lau [28] demonstrated the potential of using reinforcement learning for active defense guidance rating, although an optimal strategy was not obtained in their preliminary investigation.

It is worthy to point out that, on one hand, to better align with real-world engineering applications, research in guidance methods often needs to consider the presence of various information gaps and noise [29,30]. However, most of the existing optimal active defense guidance methods rely on perfect information assumptions, leading to subpar performance when faced with unknown prior knowledge or observation noise. Additionally, these methods often struggle to meet the real-time requirements of spacecraft applications. On the other hand, the majority of reinforcement learning algorithms have been applied to non-adversarial or weak adversarial flight missions, where mission objectives and process rewards are clear and intuitive. However, in the highly competitive TID game scenario, obtaining effective reward information becomes challenging due to the intense confrontation between agents, leading to sparse reward problems or "Plateau Phenomenon" [31].

Given these observations, there is a strong motivation to develop an active defense guidance method based on reinforcement learning that possesses enhanced real-time capabilities, adaptiveness, and robustness, while addressing the challenges posed by adversarial scenarios and sparse reward issues.

In this paper, we focus on the cooperative active defense guidance strategy design of a target spacecraft with active defense attempting to evade an interceptor in space. This TID scenario holds significant importance in the domains of space attack-defense and ballistic missile penetration. The paper begins by deriving the kinematic and first-order dynamic models of the engagement scenario. Subsequently, an intelligent cooperative active defense (ICAAI) guidance method for active defense is proposed, utilizing the twin-delay deep deterministic policy gradient (TD3) algorithm. To address the challenge of sparse rewards, an efficient and stable convergence (ESC) training approach is introduced. Furthermore, benchmark comparisons are made using Optimal Guidance Laws (OGLs), and simulation analyses are presented to validate the performance of the proposed method.

The paper is organized as follows. In Section 2, the problem formulation is provided. In Section 3, the guidance law is developed. In Section 4, experiments are presented where the proposed method has been compared with its analytical counterpart, followed by the conclusions presented in Section 5.

2. Problem Formulation

Consider a multi-agent game with a spacecraft as the main target (T), an active defense vehicle as the defender (D), and a highly maneuverable small spacecraft as the interceptor (I). In this battle, the interceptor chases the target, which launches the defender to protect itself by destroying the interceptor. During the endgame, all players are considered as constant-speed mass points whose trajectories can be linearized around the initial line of sight. As a consequence of trajectory linearization, the engagement, a three-dimensional process, can be simplified and will be analyzed in one plane. However, it should be noted that in most cases these assumptions do not affect the generality of the results [11].

A schematic view of the engagement is shown in Figure 1, where $X-O-Y$ is a Cartesian inertial reference frame. The distances between the players are denoted as ρ_{ID} and ρ_{IT}, respectively. Each player's velocity is indicated as V_{I}, V_{T}, and V_{D}, while their accelerations are represented as a_{I}, a_{T}, and a_{D}. The flight path angles of the players are defined as ϕ_{I}, ϕ_{T}, and ϕ_{D}, respectively. The line of sight (LOS) between the players is described by $\mathrm{LOS}_{\mathrm{ID}}$ and $\mathrm{LOS}_{\mathrm{IT}}$, and the angles between the LOS and the X-axis are denoted as λ_{ID} and λ_{IT}. The lateral displacements of each player relative to the X-axis are represented as y_{I}, y_{T}, and y_{D}, while the relative displacements between the players are defined as y_{IT} and y_{ID}.

Considering the collective mission objectives, the target's priority is to evade the interceptor with defender support. Simultaneously, the interceptor aims to avoid the defender while chasing the target. Consequently, the target's guidance law strives for maximum convergence, while the defender's aims for convergence to zero. Conversely, the interceptor's guidance law assumes the opposite role (as depicted in Figure 1). This

scenario can thus be segmented into two collision triangles: one involving the interceptor and the target, and the other between the interceptor and the defender.

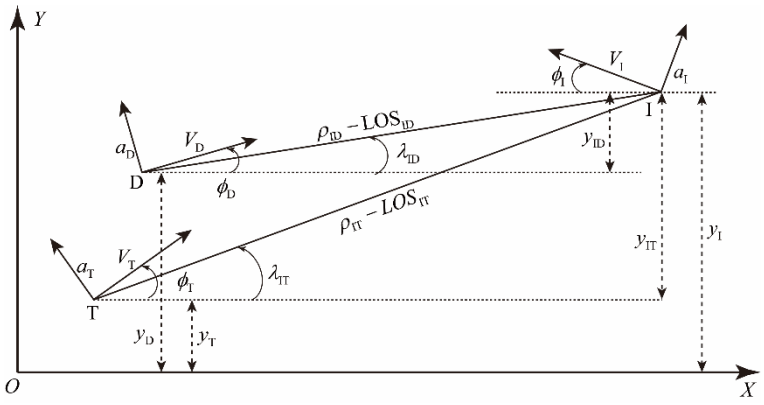


Figure 1. Schematic view of the engagement.

2.1. Equations of Motion

Consider the I-T collision triangle and the I-D collision triangle in a multi-agent pursuit-evasion engagement. The kinematics are expressed using the polar coordinate system attached in the target and defender as follows:

$$\begin{aligned} \dot{\rho}_{IT} &= -V_I \cos(\phi_I + \lambda_{IT}) - V_T \cos(\phi_T - \lambda_{IT}) \\ \dot{y}_{IT} &= V_I \sin\phi_I - V_T \sin\phi_T \\ \dot{\lambda}_{IT} &= \frac{V_I \sin(\phi_I+\lambda_{IT}) - V_T \sin(\phi_T-\lambda_{IT})}{\rho_{IT}} \end{aligned} \tag{1}$$

$$\begin{aligned} \dot{\rho}_{ID} &= -V_I \cos(\phi_I + \lambda_{ID}) - V_D \cos(\phi_D - \lambda_{ID}) \\ \dot{y}_{ID} &= V_I \sin\phi_I - V_D \sin\phi_D \\ \dot{\lambda}_{ID} &= \frac{V_I \sin(\phi_I+\lambda_{ID}) - V_D \sin(\phi_D-\lambda_{ID})}{\rho_{ID}} \end{aligned} \tag{2}$$

Furthermore, the flight path angles associated with dynamics can be defined for each of the players:

$$\dot{\phi}_i = \frac{a_i}{V_i},\ i = \{\mathrm{I},\ \mathrm{T},\ \mathrm{D}\} \tag{3}$$

2.2. Linearized Equations of Motion

In the research context, both the LOS angle λ and fight path angle ϕ are small quantities, and the inter-spacecraft distances are much larger than the spacecraft velocities. Furthermore, during the terminal guidance phase, the rate of change in spacecraft velocity magnitude approaches zero. Therefore, the equations of motion can be linearized around the initial line-of-sight:

$$\begin{aligned} \dot{\rho}_{IT} &= -V_I \cos(\phi_I + \lambda_{IT}) - V_T \cos(\phi_T - \lambda_{IT}) \approx -(V_I + V_T) \\ \ddot{y}_{IT} &= (V_I \sin\phi_I - V_T \sin\phi_T)' \approx (V_I\phi_I - V_T\phi_T)' \\ &= \dot{V}_I\phi_I - \dot{V}_T\phi_T + V_I\dot{\phi}_I - V_T\dot{\phi}_T = \dot{V}_I\phi_I - \dot{V}_T\phi_T + a_I - a_T \\ &\approx a_I - a_T \\ \dot{\lambda}_{IT} &= \frac{V_I \sin(\phi_I+\lambda_{IT}) - V_T \sin(\phi_T-\lambda_{IT})}{\rho_{IT}} \approx \frac{V_I(\phi_I+\lambda_{IT}) - V_T(\phi_T-\lambda_{IT})}{\rho_{IT}} \approx 0 \end{aligned} \tag{4}$$

$$\begin{aligned} \dot{\rho}_{ID} &= -V_I \cos(\phi_I + \lambda_{ID}) - V_D \cos(\phi_D - \lambda_{ID}) \approx -(V_I + V_D) \\ \ddot{y}_{ID} &= (V_I \sin\phi_I - V_D \sin\phi_D)' \approx (V_I\phi_I - V_D\phi_D)' \\ &= \dot{V}_I\phi_I - \dot{V}_D\phi_D + V_I\dot{\phi}_I - V_D\dot{\phi}_D = \dot{V}_I\phi_I - \dot{V}_D\phi_D + a_I - a_D \\ &\approx a_I - a_D \\ \dot{\lambda}_{ID} &= \frac{V_I \sin(\phi_I+\lambda_{ID}) - V_D \sin(\phi_D-\lambda_{ID})}{\rho_{ID}} \approx \frac{V_I(\phi_I+\lambda_{ID}) - V_D(\phi_D-\lambda_{ID})}{\rho_{ID}} \approx 0 \end{aligned} \tag{5}$$

The dynamics for each of the players is assumed to be a first-order process:

$$\dot{a}_i = -\frac{a_i - u_i}{\tau_i},\ i = \{\mathrm{I},\ \mathrm{T},\ \mathrm{D}\} \tag{6}$$

Furthermore, the variable vector can be defined as follows:

$$\boldsymbol{x} = \begin{bmatrix} y_{\mathrm{IT}} & \dot{y}_{\mathrm{IT}} & y_{\mathrm{ID}} & \dot{y}_{\mathrm{ID}} & a_{\mathrm{I}} & a_{\mathrm{T}} & a_{\mathrm{D}} \end{bmatrix} \tag{7}$$

while the linearized equations of motion in the state space form can be written as follows:

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\begin{bmatrix} u_{\mathrm{I}} & u_{\mathrm{T}} & u_{\mathrm{D}} \end{bmatrix}^T \tag{8}$$

where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1/\tau_{\mathrm{I}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1/\tau_{\mathrm{T}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1/\tau_{\mathrm{D}} \end{bmatrix} \tag{9}$$

$$\boldsymbol{B} = \begin{bmatrix} 0_{5\times 3} \\ \boldsymbol{B}_1 \end{bmatrix},\ \boldsymbol{B}_1 = \begin{bmatrix} 1/\tau_{\mathrm{I}} & 0 & 0 \\ 0 & 1/\tau_{\mathrm{T}} & 0 \\ 0 & 0 & 1/\tau_{\mathrm{D}} \end{bmatrix} \tag{10}$$

Since the velocity of each player is assumed to be constant, the engagement can be formulated as a fixed-time process. Thus, the interception time can be calculated using the following:

$$\begin{aligned} t_{\mathrm{f,IT}} &= -\rho^0_{\mathrm{IT}}/\dot{\rho}_{\mathrm{IT}} = \rho^0_{\mathrm{IT}}/(V_{\mathrm{I}} + V_{\mathrm{T}}) \\ t_{\mathrm{f,ID}} &= -\rho^0_{\mathrm{ID}}/\dot{\rho}_{\mathrm{ID}} = \rho^0_{\mathrm{ID}}/(V_{\mathrm{I}} + V_{\mathrm{D}}) \end{aligned} \tag{11}$$

where ρ^0_{IT} represents the initial relative distance between the interceptor and the target, while ρ^0_{ID} is the distance between the interceptor and the defender, allowing us to define the time-to-go of each engagement by

$$\begin{aligned} t_{\mathrm{go,IT}} &= t_{\mathrm{f,IT}} - t \\ t_{\mathrm{go,ID}} &= t_{\mathrm{f,ID}} - t \end{aligned} \tag{12}$$

which represents the expected remaining game time for the interceptor in the "Interceptor vs. Target" and "Interceptor vs. Defender" game scenarios, respectively.

2.3. Zero-Effort Miss

A well-known zero-effort miss (ZEM) is introduced in the guidance law design and reward function design. It is obtained from the homogeneous solutions of equations of motion and is only affected by the current state and interception time. It can be calculated as follows:

$$\begin{aligned} Z_{\mathrm{IT}}(t) &= \boldsymbol{L}_1\boldsymbol{\Phi}(t, t_{\mathrm{f,IT}})\boldsymbol{x}(t) \\ Z_{\mathrm{ID}}(t) &= \boldsymbol{L}_2\boldsymbol{\Phi}(t, t_{\mathrm{f,ID}})\boldsymbol{x}(t) \end{aligned} \tag{13}$$

where

$$\begin{aligned} \boldsymbol{L}_1 &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \boldsymbol{L}_2 &= \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \tag{14}$$

Thus, the ZEM and its derivative with respect to time are given as follows:

$$\begin{aligned} Z_{\mathrm{IT}}(t) &= x_1 + t_{\mathrm{goIT}}x_2 + a_{\mathrm{I}}\tau_{\mathrm{I}}^2\varphi\left(t_{\mathrm{goIT}}/\tau_{\mathrm{I}}\right)x_5 - a_{\mathrm{T}}\tau_{\mathrm{T}}^2\varphi\left(t_{\mathrm{goIT}}/\tau_{\mathrm{T}}\right)x_6 \\ Z_{\mathrm{ID}}(t) &= x_3 + t_{\mathrm{goID}}x_4 + a_{\mathrm{I}}\tau_{\mathrm{I}}^2\varphi\left(t_{\mathrm{goID}}/\tau_{\mathrm{I}}\right)x_5 - a_{\mathrm{D}}\tau_{\mathrm{D}}^2\varphi\left(t_{\mathrm{goID}}/\tau_{\mathrm{D}}\right)x_7 \end{aligned} \tag{15}$$

$$\begin{aligned}\dot{Z}_{\mathrm{IT}}(t) &= \tau_{\mathrm{I}}\varphi\left(t_{\mathrm{goIT}}/\tau_{\mathrm{I}}\right)u_{\mathrm{I}} - \tau_{\mathrm{T}}\varphi\left(t_{\mathrm{goIT}}/\tau_{\mathrm{T}}\right)u_{\mathrm{T}}\\ \dot{Z}_{\mathrm{ID}}(t) &= \tau_{\mathrm{I}}\varphi\left(t_{\mathrm{goID}}/\tau_{\mathrm{I}}\right)u_{\mathrm{I}} - \tau_{\mathrm{D}}\varphi\left(t_{\mathrm{goID}}/\tau_{\mathrm{D}}\right)u_{\mathrm{D}}\end{aligned} \tag{16}$$

where

$$\varphi(\chi) = \mathrm{e}^{-\chi} + \chi - 1 \tag{17}$$

2.4. Problem Statement

This research focuses on the terminal guidance task of evading a homing interceptor for a maneuvering target with active defense. We design a cooperative active defense guidance to facilitate coordinated maneuvers between the target and the defender based on DRL. This enables the target to evade the interceptor's interception while allowing the defender to counter-intercept the incoming threat.

3. Guidance Law Development

In this section, we develop the Intelligent Cooperative Active Defense (ICAAI) guidance strategy and design an efficient and stable convergence (ESC) training approach. The target and defender utilize ICAAI guidance, while the interceptor employs OGL. We describe the game scenario using a Markov process, present the ICAAI guidance strategy, and design an ESC training approach based on reward shaping and curriculum learning.

3.1. Markov Decision Process

The sequential decision making that an autonomous RL agent interacts with the environment (e.g., the engagement) can be formally described as an MDP, which is required to properly set up the mathematical framework of an DRL problem. A generic time-discrete MDP can be represented as a 6-tuple $\{s, o, a, P_{sa}, \gamma, R\}$. $s_t \in S \in \mathbb{R}^n$ is a vector that completely identifies the state of the system (e.g., the EOM) at time t. Generally, the complete state is not available to the agent at each time t; the decision-making relies on an observation vector $o_t \in O \in \mathbb{R}^m$. In the present paper, the observations are defined as an uncertain (e.g., imperfect and noisy) version of the true state, which can be written as a function Ω of the current state s_t. The action $a \in A \in \mathbb{R}^l$ of the agent is given by a state-feedback policy $\pi : O \to A$, that is, $a_t = \pi(o_t)$. P_{sa} is time-discrete dynamic model describing the transformation led by the state–action pair (s_t, a_t). As a result, the evolution rule of the dynamic system can be described as follows:

$$\begin{aligned}s_{t+1} &= P_{sa}(s_t, a_t)\\ o_t &= \Omega(s_t)\\ a_t &= \pi(o_t)\end{aligned} \tag{18}$$

Since a fixed-time engagement is considered, the interaction between the agent and the environment gives rise to a trajectory $\mathbf{I}$:

$$\begin{aligned}\mathbf{I} &= [\iota_1, \iota_2, \cdots, \iota_t, \cdots, \iota_{T-1}, \iota_T]\\ \iota_t &= [o_t, a_t, r_t]^T\end{aligned} \tag{19}$$

where the trajectory information at each time step ι_t is composed of observational o_t, action a_t, and reward signal r_t generated through the interaction between the agent and the environment.

The return, the agent received at time t in the trajectory $\mathbf{I}$, is defined as a discounted sum of rewards:

$$R_t^{\mathbf{I}} = \sum_{i=t}^{T} \gamma^{i-t} r_i \tag{20}$$

where $\gamma \in (0, 1]$ is a discount rate determining whether the agent has a long-term vision ($\gamma = 1$) or is short-sighted ($\gamma \ll 1$).

Prior to deriving the current guidance law, we outline the key elements of the MDP: state space, action space, and observations. We present the reward design separately by highlighting a crucial aspect of the configuration.

3.1.1. Perfect Information Model

In a deterministic model, the basic assumption is that each player has perfect information about the interceptor (e.g., states, maximum acceleration, and time constant). The communication of this information between the defender and the protected target is assumed to be ideal and without delay. Thus, the state space can be identified by states, maximum acceleration, and time constant:

$$\boldsymbol{s}_t = \begin{bmatrix} t & \boldsymbol{x}_t & \boldsymbol{y}_t & \boldsymbol{V}_t & \boldsymbol{a_t} & \boldsymbol{a}_{\max} & \boldsymbol{\tau} \end{bmatrix}^T \tag{21}$$

$$\boldsymbol{x}_t = \begin{bmatrix} x_{t,\mathrm{T}} & x_{t,\mathrm{D}} & x_{t,\mathrm{I}} \end{bmatrix}, \boldsymbol{y}_t = \begin{bmatrix} y_{t,\mathrm{T}} & y_{t,\mathrm{D}} & y_{t,\mathrm{I}} \end{bmatrix} \tag{22}$$

$$\boldsymbol{V}_t = \begin{bmatrix} V_{t,\mathrm{T}} & V_{t,\mathrm{D}} & V_{t,\mathrm{I}} \end{bmatrix} \tag{23}$$

$$\mathbf{a}_t = \begin{bmatrix} a_{t,\mathrm{T}} & a_{t,\mathrm{D}} & a_{t,\mathrm{I}} \end{bmatrix} \tag{24}$$

$$\mathbf{a}_{\max} = \begin{bmatrix} a_{\max,\mathrm{T}} & a_{\max,\mathrm{D}} & a_{\max,\mathrm{I}} \end{bmatrix} \tag{25}$$

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{\mathrm{T}} & \tau_{\mathrm{D}} & \tau_{\mathrm{I}} \end{bmatrix} \tag{26}$$

As with the multi-agent system, interactions introduce uncertainty into the environment, which significantly affects the stability of the RL algorithm. Given the full cooperation between defender and target due to communication assumptions, the model must learn a shared guidance law for both. This effectively mitigates environmental uncertainty and enhances model convergence. In practical application, the same trained agent is assigned to the target pair, yielding the following action space:

$$\text{action} = \begin{bmatrix} u_{\mathrm{T}} & u_{\mathrm{D}} \end{bmatrix} \tag{27}$$

Since the dynamics of the scenario are formulated in Section 2.1, the state can be propagated implicitly as the linearized equation of motion presented in Equations (4)–(6).

3.1.2. Imperfect Information Model

The imperfection of information is usually due to the limitations of radar measurement and the erasure of prior knowledge. However, in existing studies, perfect information is a strong assumption, which leads to implementation difficulties in practice. To address this dilemma, this thesis considers information degradation. On the one hand, the interceptor is assumed to have perfect information (i.e., the relative states and maneuverability of the target and the defender). On the other hand, the observation of the target and defender is imperfect and even noise-corrupted. The observation uncertainty is modeled as observation noise and a mask on the perfect information.

$$o_t = \Omega(s_t) = \boldsymbol{\Gamma} s_t \times (\boldsymbol{I} + \boldsymbol{\omega}_{o,t}) = \begin{bmatrix} t \\ \boldsymbol{x}_t \\ \boldsymbol{y}_t \\ \boldsymbol{V}_t \\ \boldsymbol{a}_t \end{bmatrix} + \begin{bmatrix} 0 \\ \delta \boldsymbol{x}_{o,t} \\ \delta \boldsymbol{y}_{o,t} \\ \delta \boldsymbol{V}_{o,t} \\ \delta \boldsymbol{a}_{o,t} \end{bmatrix} \tag{28}$$

where $\boldsymbol{\Gamma}$ is the mask matrix and $\boldsymbol{\omega}_{o,t}$ is the observation noise vector. $\boldsymbol{\omega}_{o,t}$ can be calculated by Equations (29)–(32).

$$\boldsymbol{\omega}_{o,t} = \begin{bmatrix} 0 \\ \delta \boldsymbol{x}_{o,t} \\ \delta \boldsymbol{y}_{o,t} \\ \delta \boldsymbol{V}_{o,t} \\ \delta \boldsymbol{a}_{o,t} \end{bmatrix} \sim U(0_{13}, \Sigma) \in \mathbb{R}^{13} \tag{29}$$

$$\Sigma = \left[0, 0, 0, \sigma_{\mathrm{xI}}, 0, 0, \sigma_{\mathrm{yI}}, 0, 0, \sigma_v, 0, 0, \sigma_a\right]^T \tag{30}$$

$$\sigma_{\mathrm{xI}} = \cos(\sigma_{\mathrm{LOS}} + \lambda_{\mathrm{IT}})\left(\rho_{\mathrm{IT}} + \sigma_\rho\right) - x_{\mathrm{I}} \approx 0 \tag{31}$$

$$\sigma_{\mathrm{yI}} = \sin(\sigma_{\mathrm{LOS}} + \lambda_{\mathrm{IT}})\left(\rho_{\mathrm{IT}} + \sigma_\rho\right) - y_{\mathrm{I}} \approx \sigma_{\mathrm{LOS}} \cdot \rho_{\mathrm{IT}} \tag{32}$$

where Σ represents the noise amplitude, with $\sigma_\rho(\mathrm{m})$, $\sigma_{LOS}(\mathrm{mrad})$, $\sigma_v(\mathrm{m/s})$, and $\sigma_a(\mathrm{m/s^2})$ the nonnegative parameters.

3.2. ICAAI Guidance Law Design

In this section, we present the mathematical framework of actor–critic RL algorithms, focusing on the algorithm used in ICAAI guidance: Twin-Delay Deep Deterministic Policy Gradient (TD3) [32]. TD3 is an advanced deterministic policy gradient reinforcement learning algorithm. In comparison to stochastic policy gradient algorithms like Proximal Policy Optimization (PPO) [33] and Asynchronous Advantage Actor–Critic (A3C) [34], TD3 exhibits a higher resistance to converging into local optima. Furthermore, when compared to traditional deterministic policy gradient RL algorithms such as Deep Deterministic Policy Gradient (DDPG) [35], TD3 achieves superior training stability and convergence efficiency. This assertion is supported by our prior RL algorithm selection experiments, as illustrated in Figure 2.

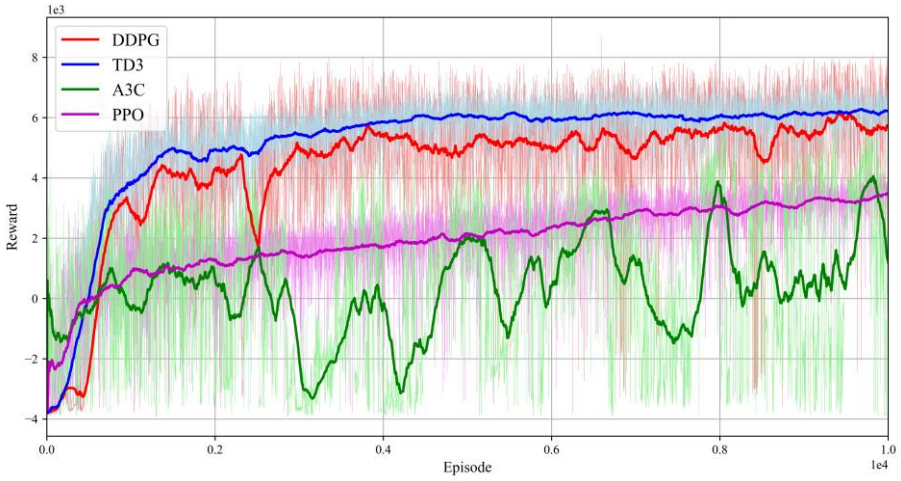


Figure 2. Comparison of training results of various reinforcement learning algorithms.

Without loss of generality, throughout the entire section the MDP is supposed to be perfectly observable (i.e., with $o_t = s_t$) to conform with the standard notation of RL. However, the perfect information state s_t can be replaced by observation o_t whenever the observations differ from the state.

3.2.1. Actor–Critic Algorithms

The RL problem's goal is to find the optimal policy π_ϕ with parameters ϕ that maximizes the expected return, which can be formulated as follows:

$$J(\phi) = \mathop{\mathbb{E}}_{\tau\sim\pi_\phi}[R_0^\tau] = \mathop{\mathbb{E}}_{\tau\sim\pi_\phi}\left[\sum_{i=0}^{T}\gamma^{i-0}r_i\right] \tag{33}$$

where $\mathop{\mathbb{E}}_{\tau\sim\pi}$ denotes the expectation taken over the trajectory τ. In actor–critic algorithms, the policy, known as the actor, can be updated by using a deterministic policy gradient algorithm [36]:

$$\nabla_\phi J(\phi) = \mathbb{E}_{P_{sa}}\left[\nabla_a Q^\pi(s,a)\Big|_{a=\pi(s)}\nabla\phi\pi\phi(s)\right] \tag{34}$$

The expected return, when performing action a in state s and following π after, is called the critic or the value function, which can be formulated as follows:

$$Q^\pi(s,a) = \mathop{\mathbb{E}}_{\tau\sim\pi_\phi}[R_t^\tau|s,a] \tag{35}$$

The value function can be learned through off-policy temporal differential learning, an update rule based on the Bellman equation which describes the relationship between the value of the state–action pair (s,a) and the value of the subsequent state–action pair (s',a'):

$$Q^\pi(s,a) = r + \gamma \mathop{\mathbb{E}}_{s',a'}[Q^\pi(s',a')] \tag{36}$$

In deep Q-learning [37], the value function can be estimated with a neural network approximator $Q_\theta(s,a)$ with parameters θ, and the network is updated by using temporary differential learning with a secondary frozen target network $Q_{\theta'}(s,a)$ to maintain a fixed objective U over multiple updates:

$$U = r + \gamma Q_{\theta'}(s',a'),\ a' = \pi_{\phi'}(s') \tag{37}$$

where the actions a' are determined by a target actor network $\pi_{\phi'}$. Generally, the loss function and update rule can be formulated as follows:

$$J(\theta) = U - Q_\theta(s,a) \tag{38}$$

$$\nabla_\theta J(\theta) = [U - Q_\theta(s,a)]\nabla_\theta Q_\theta(s,a) \tag{39}$$

The parameters of target networks are updated periodically to exactly match the parameters of the corresponding current networks, which is called delayed update. This leads to the original actor–critic method, the basic structure of which is shown in Figure 3.

3.2.2. Twin-Delayed Deep Deterministic Policy Gradient Algorithm

To address the common RL issues in actor-critic algorithms (i.e., overestimation bias and accumulation of errors), in the TD3 algorithm, the actor–critic framework is modified from three aspects.

A novel variant of double Q-learning [38] called clipped double Q-learning is developed to limit possible overestimation. This provides the update objective of the critic:

$$U = r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \pi_{\phi'_1}(s')) \tag{40}$$

The parameters of policy networks are updated periodically to match the value network, which is called delayed policy update, and the soft update approach is adopted, which can be formulated as follows:

$$\theta' \leftarrow \kappa\theta + (1-\kappa)\theta' \tag{41}$$

where κ is a proportion parameter.

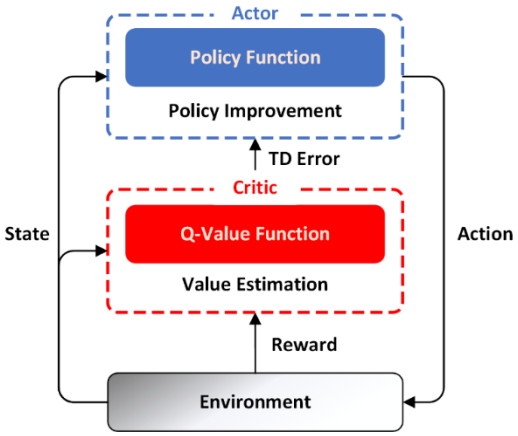


Figure 3. Structure of actor–critic method.

Target policy smoothing regularization is adopted to alleviate the overfitting phenomenon, which can be explicated as follows:

$$y = r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s') + \varepsilon) \tag{42}$$

where ε is a clipped Gaussian noise.

An overview of the TD3 algorithm is demonstrated in Figure 4.

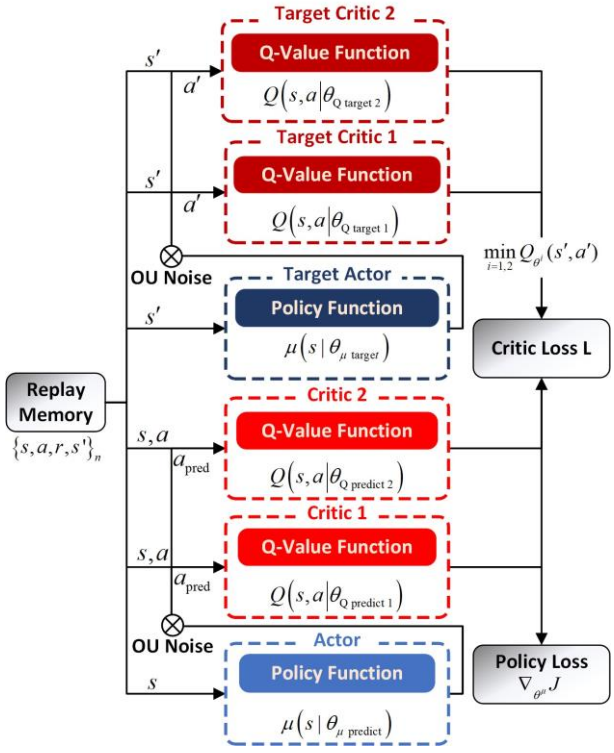


Figure 4. Structure of the TD3 algorithm.

3.2.3. Implementation Details

As for the network architecture setting, the agent observations are vectors with 13 dimensions. Both the guidance policy estimation (actor) and the value function estimation (critic) consist of three fully connected layers with sizes of 64, 256, and 512, respectively, along with layer normalization. The output layer has two units for the actor, representing the unified command of the target and defender, respectively, and one unit for the critic. The activation function is ReLU for the hidden layer neurons and linear for the output layer neuron. This structure is heuristically designed and can be generalized for efficient function approximation. Deeper and wider networks are avoided for real-time performance and fast convergence.

The hyperparameters of TD3 have been devised and validated by empirical experiments, which are reported in Table 1.

Table 1. TD3 hyperparameters.

Hyperparameter	Symbol	Value
Discount factor	γ	0.99
Learning rate	α	3×10^{-4}
Buffer size	$\mathbb{B}$	5120
Batch size	n_{batch}	128
Soft update coefficient	ζ	5×10^{-3}
Policy delay	n_{opt}	2
Train frequency	ω	6000

3.3. ESC Traning Technique

Aiming at the sparse reward problem in the multi-agent pursuit-evasion game, an efficient and stable convergence (ESC) training approach of reinforcement learning is proposed based on reward shaping [39] and curriculum learning [40].

3.3.1. Reward Shaping

The design of a reward function is the most challenging part of solving this multi-agent pursuit-evasion game through RL, as the function had to be adaptive to engagement with a sparse reward setting. It is found that, except for the common leadership mission, the pursuit-evasion game can be formulated as a strictly competitive zero-sum game. In addition, the agent policy network weights were randomly initiated at the beginning of training, while the interceptor was deployed with optimal guidance and is sufficiently aggressive.

In [41], a shaping technique was presented as a particularly effective approach to solving sparse reward problems through a series of biological experiments. The researchers divided a difficult task into several simple units and trained the animals according to an easy-to-hard schedule. This approach requires adjusting the reward signal to cover the entire training process, followed by gradual changes in task dynamics as training progresses. In [40], researchers took this idea further and proposed curriculum learning, a type of training strategy. In this work, the shaping technique and curriculum learning were used to speed up the convergence of neural networks and to increase the stability and performance of the algorithm.

The goal of the target and the defender is to converge Z_{ID} to zero as $t \to t_{f2}$ while keeping Z_{IT} as large as possible. On the contrary, the interceptor control law is designed to make Z_{IT} converge to zero while maintaining Z_{ID} as large as possible.

For this reason, a non-sparse reward function is defined in Equations (43) and(44):

$$r_{\text{medium}} = \gamma \Phi\left(s'\right) - \Phi(s) \tag{43}$$

$$\Phi(s) = \left|\frac{Z_{\mathrm{IT}}}{\alpha_1}\right|^{\beta_1} - \left|\frac{Z_{\mathrm{ID}}}{\alpha_2}\right|^{\beta_2} \tag{44}$$

$$r_{\mathrm{terminal}} = \begin{cases} \sigma, \text{ if succeed} \\ -\sigma, \text{ else} \end{cases} \tag{45}$$

where γ is the discount factor in the Markov decision process and α_1, α_2, β_1, β_2, and σ are the positive hyperparameters.

It must be stressed that, since both the number and maneuverability of players completely change the environment, the hyperparameter values used in this paper may be not universal. Thus, in the following subsection, the focus will be on the applied design method instead of the specific hyperparameter values. The r_{terminal} is the terminal reward signal given to the terminal behavior of the agent, which is sparse but intuitive. Situations in which the interceptor is destroyed by the defender (when $t = t_{\mathrm{f,ID}}$) or when the interceptor is driven away by the defender and misses the target are judged as a success. Furthermore, the r_{medium} is a non-sparse reward function based on the difference form of a potential function $\Phi(s)$ which ensures the consistency of the optimal strategy [42–44]. It is important to emphasize that the design of $\Phi(s)$ relies on a fractional exponential function. This function provides a continuous reward signal for the agent's evaluation of each state. Notably, this model exhibits a unique property: as the base number approaches infinity, the gradient decreases to zero, and as the base number approaches zero, the gradient increases infinitely. This specific characteristic significantly aids the agent in converging towards states where the base number is either greater than zero or approaches zero.

In this paper, the defined reward function carries the physical meaning of the mission—the target must escape from the interceptor, while the defender has to get close to the interceptor. The r_{medium} value increases as Z_{ID} converges to zero, or when Z_{IT} increases. On the other hand, it decreases when Z_{ID} is divergent or when Z_{IT} converges to zero.

Generally, reward normalization is beneficial to neural network convergence. However, determining the bounds of Z_{IT} and Z_{ID} is a complex task. For this reason, hyperparameters α_1, α_2, β_1, and β_2 are tuned, aiming to scale the r_{medium} close to $[-c,\ c]$, in which c is a positive constant. In the following step, the design of ρ is considered, which introduces the expectation of agent foresight. If the agent is expected to predict the terminal reward r_{terminal} n steps before, the discounted terminal reward must be larger than the r_{medium} bounds. Thus, the hyperparameter ρ satisfies the following expression:

$$\rho \geq \frac{c}{\gamma^n} \tag{46}$$

3.3.2. Curriculum Learning

After hyperparameter tuning, we enhance the training stability of intelligent algorithms using an adaptive progressive curriculum learning approach. This method incrementally raises training complexity to enhance agent capability and performance. The agent's training level is adaptively assessed through changes in network loss, determining appropriate training difficulty. The v_{PG} calculation formula is as follows:

$$v_{\mathrm{PG}} = L(x, \theta) - L\left(x, \theta'\right) \tag{47}$$

where $L(\cdot)$ represents the calculation function of network loss; θ is the current network parameter and θ' is the new network parameter obtained after data x training. Given a small amount $\varepsilon(1 \gg \varepsilon > 0)$, when

$$|v_{\mathrm{PG}}| < \varepsilon \tag{48}$$

the agent training enters the next stage. A sequence of increasingly difficult tasks is allocated to the agent, as shown in Table 2. The curriculum was divided into three stages:

- The agent is required to combat the interceptors employing non-maneuvering;

- Square wave signal;
- OGL.

Finally, it is possible to complete the reward shaping process.

Table 2. Curriculum learning.

Curriculum	Stage 1	Stage 2	Stage 3
Interceptor guidance command	None	Square wave signal	OGL
Maximum interceptor acceleration	0	8 g	4 g/6 g/8 g

In summary, the block diagram of ICAAI guidance strategy is shown in Figure 5.

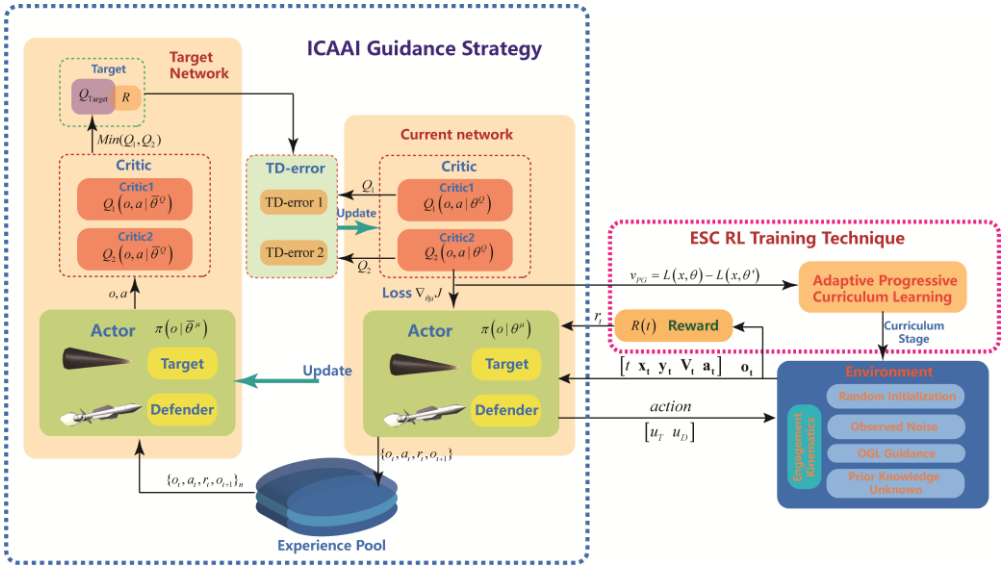


Figure 5. Block Diagram of ICAAI Guidance Strategy.

4. Experiments

In this section, we demonstrate the efficacy of the proposed guidance method and the effectiveness of the shaping technique through learning processes and Monte Carlo simulations. We establish benchmark comparisons by including OGLs and evaluating application requirements. To illustrate, we consider a scenario [10] involving a maneuverable small spacecraft (Interceptor, I), a defensive vehicle (Defender, D), and an evading spacecraft (Target, T), all in circular Earth orbits. Gravity effects are incorporated in the simulations. Assumptions include the interceptor's superior maneuverability and time constant compared to the target and defender.

4.1. Optimal Pursuit and Evasion Guidance Laws

Lemma 1. *The linear–quadratic optimal guidance law (LQOGL) [10]:*

$$u_I^* = \begin{cases} -\frac{K(t)Z_{\mathrm{ID}}(t)}{\omega_1} u_{\mathrm{I}}^{\max} \tau_{\mathrm{I}} \varphi\left(\frac{t_{\mathrm{f,ID}}-t}{\tau_{\mathrm{I}}}\right) & \text{for } \|Z_{\mathrm{ID}}(t)\| < \eta \\ -\frac{P(t)Z_{\mathrm{IM}}(t)}{\xi_1} u_{\mathrm{I}}^{\max} \tau_{\mathrm{I}} \varphi\left(\frac{t_{\mathrm{f,IT}}-t}{\tau_{\mathrm{I}}}\right) & \text{else} \end{cases} \tag{49}$$

where η is a positive constant representing the limit-collision radius between the interceptor and the defender, and $u_{\mathrm{I}}^{\max}$ is the maximum control force provided by the interceptor. Furthermore, variable K(t) and P(t) can be defined as follows:

$$K(t) = \frac{1}{\int_t^{t_{\mathrm{fID}}} \left[\frac{1}{\omega_1} \left(u_{\mathrm{I}}^{\max} \tau_{\mathrm{I}} \varphi\left(\frac{t_{\mathrm{f,ID}} - t}{\tau_1} \right) \right)^2 - \frac{1}{\omega_2} \left(u_D^{\max} \tau_{\mathrm{D}} \varphi\left(\frac{t_{\mathrm{f,ID}} - t}{\tau_{\mathrm{D}}} \right) \right)^2 \right] \mathrm{d}t - 1} \tag{50}$$

$$P(t) = \frac{1}{\int_t^{t_{\mathrm{fIM}}} \left[\frac{1}{\xi_1} \left(u_{\mathrm{I}}^{\max} \tau_{\mathrm{I}} \varphi\left(\frac{t_{\mathrm{f,IM}} - t}{\tau_1} \right) \right)^2 - \frac{1}{\xi_2} \left(u_{\mathrm{M}}^{\max} \tau_{\mathrm{M}} \varphi\left(\frac{t_{\mathrm{f,IM}} - t}{\tau_{\mathrm{M}}} \right) \right)^2 \right] \mathrm{d}t - 1} \tag{51}$$

whereω_1, ω_2, ξ_1, and ξ_2 are nonnegative constants ensuring the interceptor converges towards the target, guaranteeing its escape from the defender.

Proof. The detailed proof of similar results can be found in [10]; see Theorem 1 and the associated proof. □

Lemma 2. *Standard optimal guidance law (SOGL) [45]:*

$$\begin{aligned} u_{\mathrm{I}}^* &= u_{\mathrm{I}}^{\max} \mathrm{sgn}[Z_{\mathrm{ID}}(t_{\mathrm{f,ID}})] \mathrm{sgn}\left[\varphi\left(\frac{t_{\mathrm{f,ID}} - t}{\tau_{\mathrm{I}}} \right) \right] \text{ for } \|Z_{\mathrm{ID}}(t)\| < \eta \\ u_{\mathrm{I}}^* &= -u_{\mathrm{I}}^{\max} \mathrm{sgn}[Z_{\mathrm{IT}}(t_{\mathrm{f,IT}})] \mathrm{sgn}\left[\varphi\left(\frac{t_{\mathrm{f,IT}} - t}{\tau_{\mathrm{I}}} \right) \right] \text{ else} \end{aligned} \tag{52}$$

whereη is a positive constant representing the switching condition always equal to the defender kill radius.

Proof. Consider the following cost function:

$$\begin{aligned} J_1 &= -\tfrac{1}{2} Z_{\mathrm{ID}}^2(t_{\mathrm{fID}}) \text{ for } \|Z_{\mathrm{ID}}(t)\| < \eta \\ J_2 &= \tfrac{1}{2} Z_{\mathrm{IT}}^2(t_{\mathrm{fIT}}) \quad \text{else} \end{aligned} \tag{53}$$

For J_1, the Hamiltonian of the problem is defined as follows:

$$H_1 = \lambda_1 \dot{Z}_{\mathrm{ID}}(t) \tag{54}$$

The costate equation and transversality condition are provided by the following:

$$\dot{\lambda}_1(t) = -\frac{\partial H_1}{\partial Z_{\mathrm{ID}}} = 0 \tag{55}$$

$$\lambda_1(t_{\mathrm{fID}}) = \frac{\partial J_1}{\partial Z_{\mathrm{ID}}(t_{\mathrm{fID}})} = -Z_{\mathrm{ID}}(t_{\mathrm{fID}}) \tag{56}$$

The optimal interceptor controller minimizes the Hamiltonian satisfying the following:

$$u_{\mathrm{I}}^* = \underset{u_{\mathrm{I}}}{\mathrm{argmin}}(H_1) \tag{57}$$

The interceptor guidance law can thus be obtained:

$$u_{\mathrm{I}}^* = u_{\mathrm{I}}^{\max} \mathrm{sgn}[Z_{\mathrm{ID}}(t_{\mathrm{fID}})] \mathrm{sgn}\left[\varphi\left(\frac{t_{\mathrm{fID}} - t}{\tau_{\mathrm{I}}} \right) \right] \tag{58}$$

For J_2, a similar interceptor guidance law can be found:

$$u_{\mathrm{I}}^* = -u_{\mathrm{I}}^{\max} \mathrm{sgn}\left[Z_{\mathrm{IT}}\left(t_{f\mathrm{IT}} \right) \right] \mathrm{sgn}\left[\varphi\left(\frac{t_{f\mathrm{IT}} - t}{\tau_{\mathrm{I}}} \right) \right] \tag{59}$$

Finally, the interceptor guidance schemes for evading the defender and pursuing the target are proposed after combining Equations (58) and (59):

$$\begin{aligned} u_{\mathrm{I}}^{*} &= u_{\mathrm{I}}^{\max}\mathrm{sgn}\left[Z_{\mathrm{ID}}\left(t_{f\mathrm{ID}}\right)\right]\mathrm{sgn}\left[\varphi\left(\frac{t_{f\mathrm{ID}}-t}{\tau_{\mathrm{I}}}\right)\right] \text{ for } \|Z_{\mathrm{ID}}(t)\| < \eta \\ u_{\mathrm{I}}^{*} &= -u_{\mathrm{I}}^{\max}\mathrm{sgn}\left[Z_{\mathrm{IT}}\left(t_{f\mathrm{IT}}\right)\right]\mathrm{sgn}\left[\varphi\left(\frac{t_{f\mathrm{IT}}-t}{\tau_{\mathrm{I}}}\right)\right] \text{ else} \end{aligned} \tag{60}$$

□

4.2. Engagement Setup

In this scenario, a target carrying an active anti-interceptor is threatened by a KKV interceptor in orbit at an altitude of 500 km. The defender maintains an initial safe distance of approximately 50 m longitudinally and 10 km transversely to the target. Given that the detection range of the interceptor's guided warhead is about 100 km, the initial transverse distance between the interceptor and the target is set at 100 km, and the initial longitudinal position is random in the range 499.8–500.2 km. In addition, the maneuverability and control response speed of the interceptor are better than those of the target and defender, and the OGL is used for guidance.

The comprehensive list of engagement parameters is shown in Table 3.

Table 3. Engagement parameters.

Parameters \ Interceptor	Interceptor	Target	Defender
Horizonal location (km)	100	0	0~15
Vertical location (km)	499.8~500.2	500	500.05
Horizonal velocity (km/s)	−3	2	2
Vertical velocity	0	0	0
Maximum acceleration (g)	8	2	6
Time constant (s)	0.02	0.1	0.05
Kill radius (m)	0.25	0.5	0.15

Furthermore, Gaussian noise with standard variance of $\sigma_{LOS} = 1$ mrad, $\sigma_v = 0.2$ m/s, and $\sigma_a = 1$ m/s^2 is considered in the interceptor information obtained by the target and defender through a radar seeker.

4.3. Experiment 1: Real-Time Performance of the Guidance Policy

To verify that the proposed RL training approach ESC can improve convergence efficiency and stability, the learning processes were demonstrated using the sparse reward (SR) signal and ESC, respectively, with the same hyperparameters. During the learning process, the weights of the neural network model were stored every 100 episodes for subsequent analysis. In addition, to remove stochasticity as a confounding factor, six random seeds were set for each case. Meanwhile, the real-time performance of the optimized agent is evaluated by comparing it with the traditional OGLs.

The agents were obtained after a training of 20,000 episodes, which took 12 h with 8 parallel workers on a computer equipped with a 104-core Intel Core Xeon Platinum 8270 CPU @2.70 GHz. Similarly, both the traditional methods and the proposed method are provided a current state or observation and return the required action. Table 4 shows the comparison of computational cost and update frequency obtained by using SOGL, LQOGL, and the proposed method. It can be seen from the table that LQOGL is time-consuming due to the calculation of the Riccati function, which is the reason why it has not been applied in practice. As a proven approach, the SOGL has excellent real-time performance. The proposed method achieved an update frequency of 10^3 Hz and showed great potential for on-board applications. While a variety of approaches (e.g., pruning and distillation) were effective to compress the policy network and further improve its real-time performance, it is not the main work of this research.

Table 4. Statistics of time consumption with different guidance methods.

Metrics	LQOGL	SOGL	ICAAI
Duration (1e3 step)	2.773 s	0.0145 s	0.910 s
Update frequency	≈360 HZ	$\approx 6.9 \times 10^4$ HZ	$\approx 1.1 \times 10^3$ HZ

Remark 1. *As shown in Equations (18) and (19), the LQOGL has to solve the Riccati differential equation. However, the experimental results show that its update frequency cannot meet the real-time requirements of spacecraft guidance. Compared to the LQOGL, the SOGL in Equation (60) does not need to solve the Riccati differential equation and has no hyperparameter. This improves both its computational efficiency and robustness at the cost of flexibility and the occurrence of the chattering phenomenon. To take into account the practical situation, the SOGL was chosen as an OGL benchmark.*

4.4. Experiment 2: Convergence and Performance of the Guidance Policy

The performance of the trained agent in the fully observable game was investigated by comparing the escape success rate corresponding to an optimized policy $\pi_\phi(s)$, obtained by performing Monte Carlo simulation in the fully observable (deterministic and with default engagement parameters) environment, with the solution of the SOGL.

4.4.1. Baselines

The SOGL for the target and the defender were considered as an OGL benchmark. Through a brief derivation similar to that in Section 3, it can be proven that the SOGLs for the target and the defender are as follows:

$$\begin{aligned} u_{\mathrm{T}} &= -u_{\mathrm{T}}^{\max} \operatorname{sgn}\left[Z_{\mathrm{IT}}\left(t_{f\mathrm{IT}}\right)\right] \operatorname{sgn}\left[\varphi\left(\frac{t_{f\mathrm{IT}}-t}{\tau_{\mathrm{T}}}\right)\right] \\ u_{\mathrm{D}} &= u_{\mathrm{D}}^{\max} \operatorname{sgn}\left[Z_{\mathrm{ID}}\left(t_{f\mathrm{ID}}\right)\right] \operatorname{sgn}\left[\varphi\left(\frac{t_{f\mathrm{ID}}-t}{\tau_{\mathrm{D}}}\right)\right] \end{aligned} \tag{61}$$

4.4.2. Convergence and Escape Success Rate

Figure 6 displays the learning curves depicting the mean accumulated reward across learning episodes for various scenarios. As depicted, in the ESC case, the agent's reward consistently escalated throughout the training episodes, ultimately stabilizing at around 6000 after 4000 iterations. Conversely, within the sparse reward (SR) framework, the ICAAI encountered a plateau phenomenon during training, resulting in an unstable convergence process for the associated reward function and eventual convergence failure.

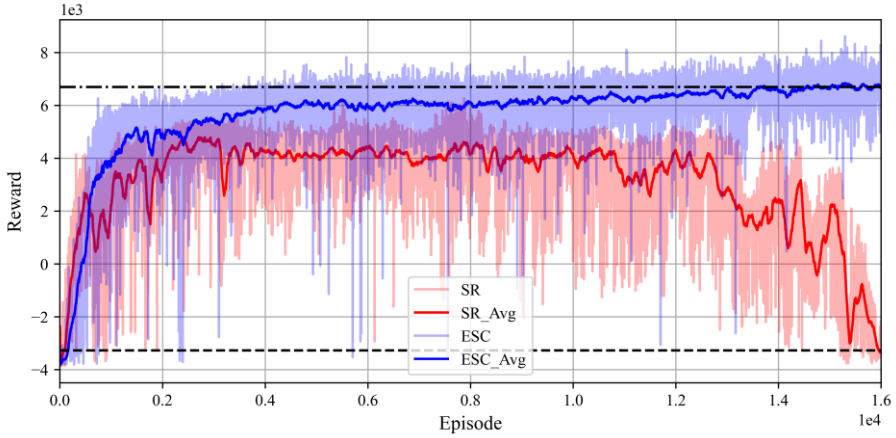


Figure 6. Learning curves of the ICAAI.

Figure 7 presents success rate curves for target evasion over learning episodes, comparing agents trained with and without ESC. The green line denotes OGL's deterministic

environment success rate of 83.4%. The ESC-trained agent surpassed the baseline by 2700 episodes, achieving a peak performance of 99% after around 13,800 episodes. Conversely, the agent without ESC exhibited a gradual decline in performance after reaching a zenith of 77%, signifying policy network overfitting during continued training. The ESC-trained agent demonstrated accelerated convergence and improved local optima. It can be inferred that the proposed ESC training approach effectively organizes exploration, addressing sparse reward issues and showcasing heightened learning efficiency and asymptotic performance. Furthermore, the proposed methodology adeptly mitigates overfitting phenomena.

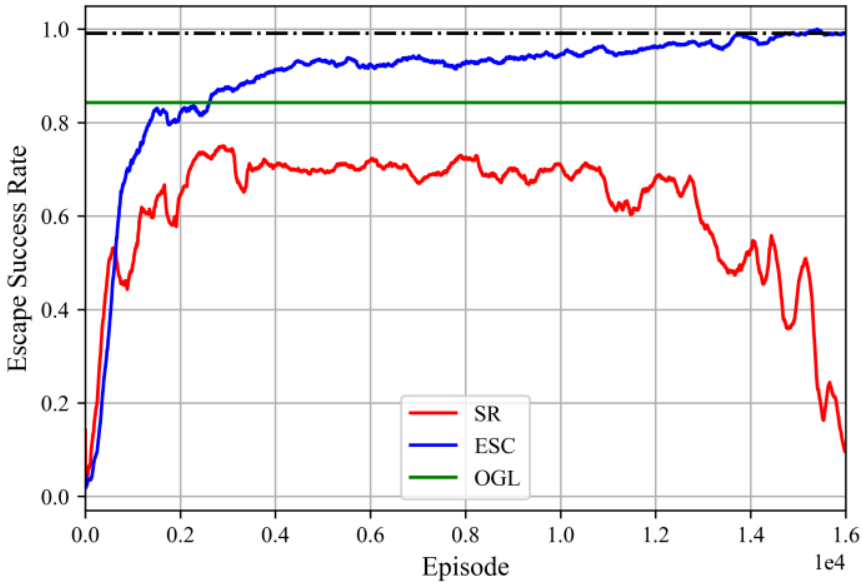


Figure 7. Escape success rate.

4.4.3. Performance Test

Figure 8 depicts spacecraft trajectories, featuring the interceptor's actual path (blue curve) and the observed trajectory from the target's perspective (yellow curve). Figure 9 displays the lateral acceleration profiles for each spacecraft, while Figure 10 illustrates the ZEM measurements between the target and interceptor and between the defender and interceptor. The simulation results presented in Figure 11 reaffirm the impact of the relative distance between the target and defender dis_{DT} on the game outcomes for the target.

Figures 8–10 illustrate the evident cooperation between the target and the defender, utilizing relative state information. Taking the simulation results at $dis_{\mathrm{DT}} = 10$ km as an example, the miss distance between the target and the interceptor was approximately 15 m. The defender maintained a miss distance of less than 1 m from the interceptor, confirming its successful interception threat. Figures 9 and 10 depict that, within 16 s of the scenario's initiation, the target collaborated with the defender, executing subtle maneuvers to intercept the interceptor. At around the 16 s mark, the interceptor perceived the threat and initiated an escape strategy. Simultaneously, the target executed an evasive maneuver in the opposite direction, utilizing its maximum maneuverability, which resulted in an increase in distance. Ultimately, the interceptor managed to evade the defender's interception attempt but failed to intercept the target in time, leading to the target's successful evasion.

In addition, the above simulation results show that the relative distance between the target and defender dis_{DT} directly determines the time it takes for the interceptor to intercept the target after evading the defender. Consequently, dis_{DT} significantly influences the game outcomes for the target, including the success rate of evasion and miss distance. Therefore, to explore the effect of dis_{DT} on the performance of ICAAI, the game results for dis_{DT} ranging from 0 to 15 km are introduced in Figure 11.

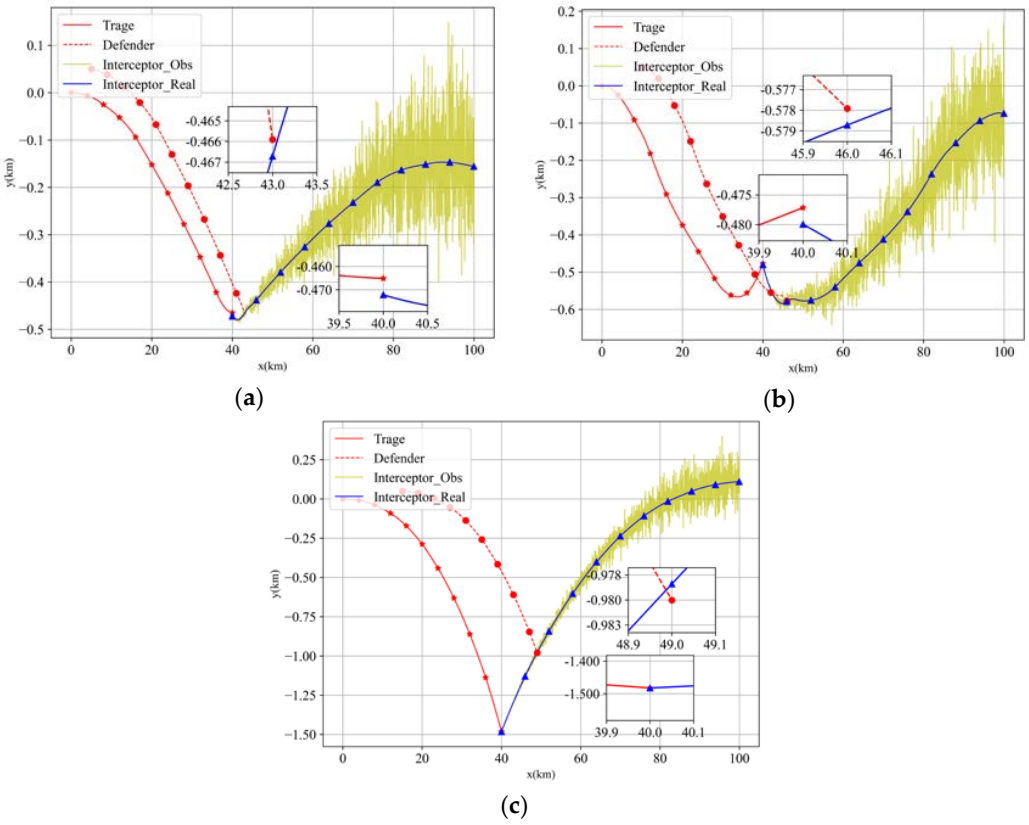


Figure 8. Spacecrafts game trajectory. (**a**) $dis_{\mathrm{DT}} = 5$ km, (**b**) $dis_{\mathrm{DT}} = 10$ km, (**c**) $dis_{\mathrm{DT}} = 15$ km.

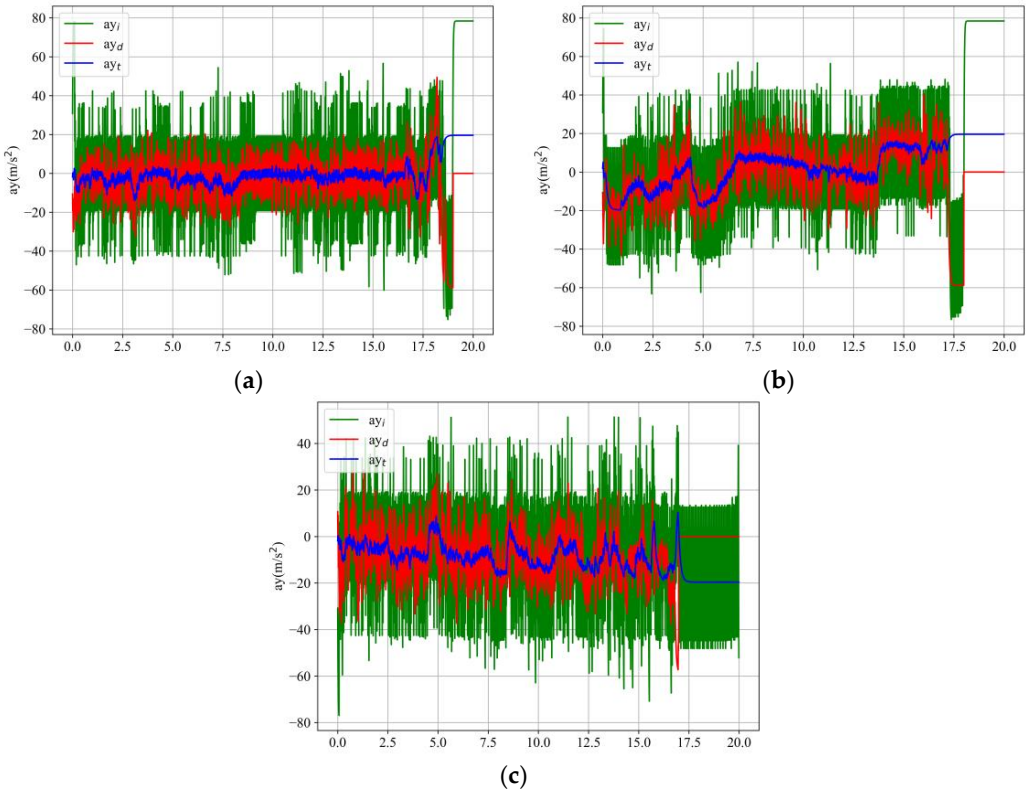


Figure 9. Lateral acceleration curve of each spacecraft. (**a**) $dis_{\mathrm{DT}} = 5$ km, (**b**) $dis_{\mathrm{DT}} = 10$ km, (**c**) $dis_{\mathrm{DT}} = 15$ km.

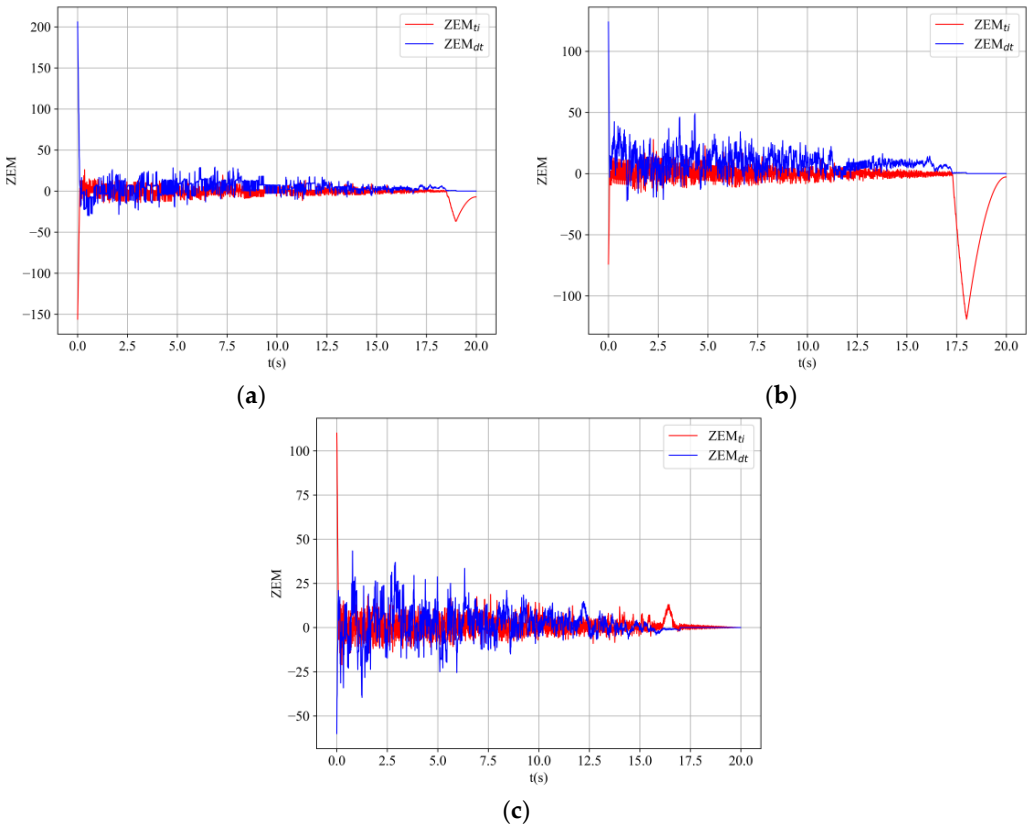


Figure 10. ZEM curve between each spacecraft. (a) dis_{DT} = 5 km, (b) dis_{DT} = 10 km, (c) dis_{DT} = 15 km.

As evident from Figure 11, employing the ICAAI intelligent game algorithm results in the target achieving success rates of no less than approximately 90% when the relative distance to the defender is less than 10 km. However, as the dis_{DT} increases from 10 to 15 km, the success rate of target evasion decreases from 90% to 0%. These simulation results illustrate that a smaller relative distance leads to an increased evasion success rate. Additionally, the curve depicting the average miss distance for the target reveals that the miss distance follows a pattern of initially increasing and then decreasing with dis_{DT}. The miss distance reaches its maximum value of approximately 50 m around a relative distance of 5 km. The occurrence of this phenomenon can be attributed to the fact that, when dis_{DT} is less than 5 km, the miss distance increases with the target's evasion time. Moreover, at this point, the interceptor has not had sufficient time to alter its trajectory to intercept the target. Conversely, when dis_{DT} exceeds 5 km, the interceptor has ample time to intercept the target after evading the defender. Consequently, the miss distance decreases with an increasing dis_{DT}.

4.5. Experiment 3: Adaptiveness of the ICAAI Guidance

In the real-world game confrontation process, obtaining the opponent's prior knowledge, such as the maximum acceleration and time constant, is often impractical. To assess the proposed ICAAI guidance method's superior adaptability compared to the OGL method under conditions of unknown opponent knowledge, several comparison conditions were designed and evaluated using the Monte Carlo target shooting method. The adaptive capabilities of both methods were analyzed based on the game results (escape success rate and miss distance) of the target spacecraft employing the two strategies.

While the target utilized OGL guidance, we considered it adopting $\overline{u}_I^{\max} = 8$ g, $\overline{\tau}_I = 0.02$ s as the prediction of the prior knowledge of the interceptor, while the actual $u_I^{\max} = 6 \sim 10$ g, $\tau = 0.05 \sim 0.002$ s. The simulation results are shown in Figure 12.

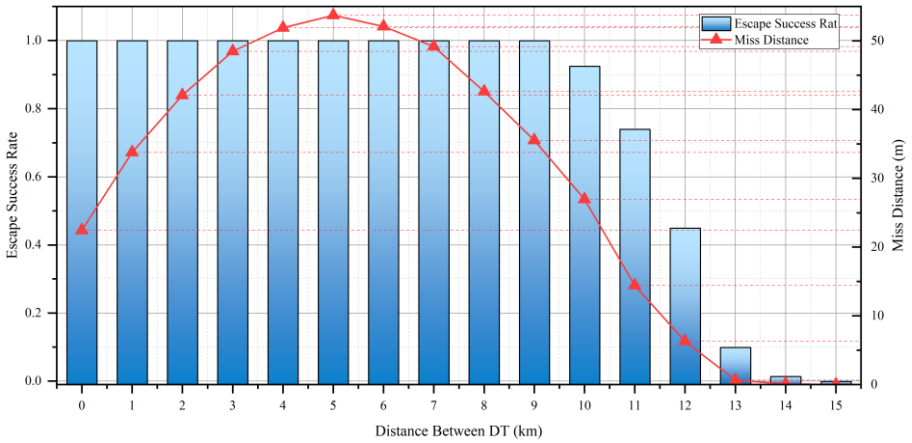
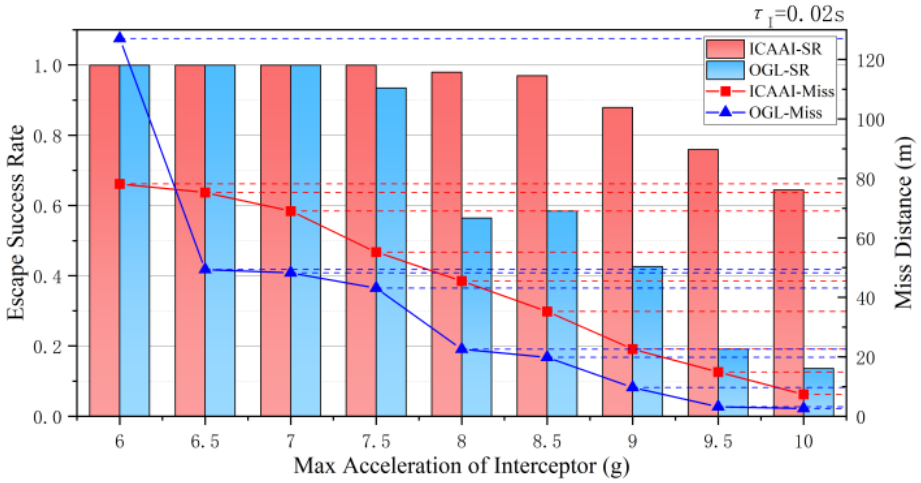
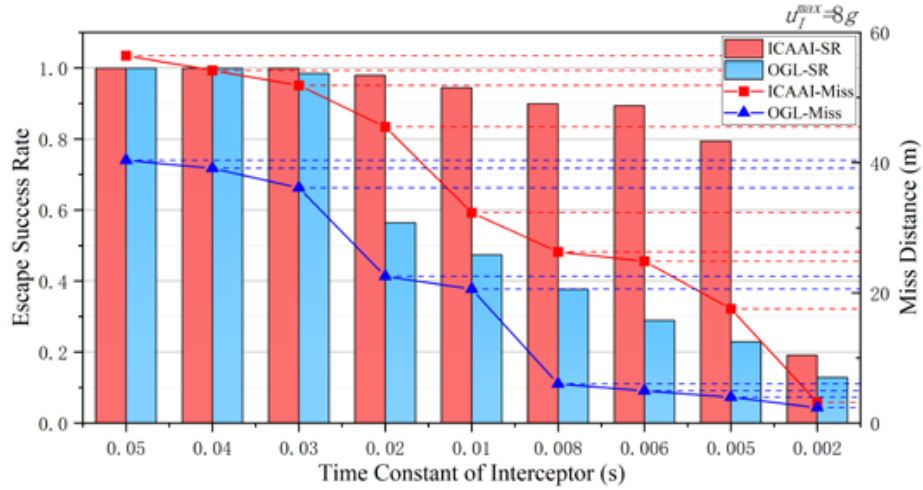


Figure 11. Target game results under different distances between target and defender.



(**a**)



(**b**)

Figure 12. Simulation results in situations without prior knowledge. (**a**) $u_I^{\max} = 6 \sim 10$ g, $\tau = 0.02$ s, (**b**) $u_I^{\max} = 8$ g, $\tau = 0.05 \sim 0.002$ s.

As depicted in Figure 12a, as the interceptor's maneuverability improves, the target's escape ability decreases for both guidance methods. However, it is evident that, when employing the ICAAI guidance, the rate of decline in the target's escape ability is significantly

lower compared to the OGL guidance method. Similarly, Figure 12b demonstrates that an increase in the interceptor's response speed yields a similar trend in the target's escape ability as in Figure 12a. Specifically, when accurately estimating the prior knowledge of the target, the escape abilities of both methods are comparable. However, when the prior knowledge error exceeds 25%, the OGL guidance leads to a reduction of over 75% in the target's escape ability, while the ICAAI guidance results in less than a 34% decrease. In conclusion, the proposed ICAAI guidance exhibits superior adaptability compared to the OGL guidance when the interceptor's prior knowledge is unknown.

Remark 2. *As an analytical method, the SOGL is stable but inflexible due to its theoretical framework [46] and stringent assumptions [47]. Correspondingly, the ICAAI control strategies are flexible and can be continuously optimized. The proposed method is independent of the time constant, which means that it performs better with less prior knowledge than the OGL. Furthermore, the adaptability of the proposed method can be improved by considering the tolerance of the maximum interceptor acceleration.*

4.6. Experiment 4: Robustness of the RL-Based Guidance Method

In addition to the unperturbed, fully observable game, the following noisy, partially observable game studies have been analyzed separately in this manuscript. The parameters used to describe the imperfect information model defined in Section 3 are shown in Table 5. The Monte Carlo simulation method is used to obtain the escape success rate and the miss distance of the target using the proposed ICAAI guidance and SOGL guidance under different noise conditions. The results of the Monte Carlo simulation are shown in Figure 13.

Table 5. Parameters of the different imperfect information models.

Measurement Noise	Parameter	Case 1	Case 2	Case 3
LOS	σ_{LOS}(mrad)	0.05	0~0.2	0.05
Velocity	σ_v(m/s)	0.2	0.2	0~0.5
Acceleration	σ_a(m/s)	1~3	2	2

Based on the simulation results of Case 2, it was observed that the OGL method exhibited significant sensitivity to LOS noise. In scenarios without LOS noise, the escape success rate of the proposed ICAAI guidance matched that of the OGL guidance, and, in some cases, the OGL method even achieved a larger miss distance. However, as the LOS noise variance increased to 0.05 mrad, the success rate of the OGL method dropped to approximately 50%. Eventually, at a LOS noise variance of 0.15 mrad, the target was practically unable to escape using the SOGL method, while the ICAAI guidance still maintained an escape success rate of around 80%.

Analyzing the simulation results of Case 1 and Case 3, it was found that due to the presence of LOS noise, the target employing the OGL method exhibited reduced sensitivity to acceleration and velocity noise. Nevertheless, its escape capability remained weaker compared to that of the ICAAI guidance. This could be attributed to the policy network propagating observation information with different weights, leveraging the exploration mechanism of reinforcement learning (RL). Consequently, training the agent in a deterministic environment resulted in a robust guidance policy with strong noise-resistant ability.

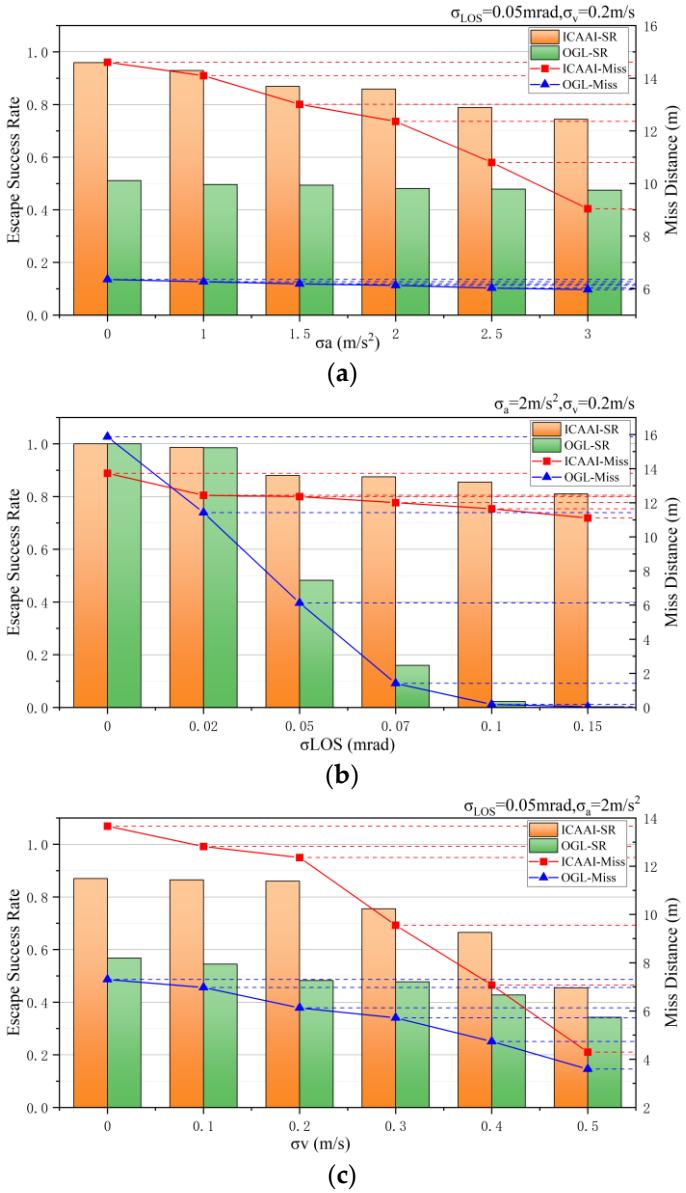


Figure 13. Simulation results in noise-corrupted environment. (**a**) Case 1, (**b**) Case 2, (**c**) Case 3.

5. Conclusions

In this research, we solved the cooperative active defense guidance problem for a target with active defense attempting to evade an interceptor. Based on deep reinforcement learning algorithms, a collaborative guidance strategy termed ICAAI was formulated to enhance active spacecraft defense. Monte Carlo simulations were conducted to empirically substantiate the real-time performance, convergence, adaptiveness, and robustness of the introduced guidance strategy. The conclusions are stated as follows:

(1) In the presence of less prior knowledge and observation noise, the proposed ICAAI guidance strategy is effective in achieving a higher success rate of target evasion by guiding the target to coordinate maneuvers with defensive spacecraft.

(2) Utilizing a heuristic continuous reward function and an adaptive progressive curriculum learning method, we devised the ESC training approach to effectively tackle issues of low convergence efficiency and training process instability in ICAAI.

(3) The ICAAI guidance strategy outperforms the linear–quadratic optimal guidance law (LQOGL) [10] in real-time performance. This framework also achieved an impressive

update frequency of 10^3 Hz, demonstrating substantial potential for onboard applications.

(4) Simulation results confirm ICAAI's effectiveness in reducing the relative distance between interceptor and defender, enabling successful target evasion. In contrast to traditional OGL methods, our approach exhibits enhanced robustness in noisy environments, particularly in mitigating line-of-sight (LOS) noise.

Author Contributions: Conceptualization, W.N., J.L., Z.L., P.L. and H.L.; Methodology, W.N., J.L., Z.L., P.L. and H.L.; Software, W.N.; Validation, W.N.; Investigation, W.N.; Data curation, W.N.; Writing—original draft, W.N.; Writing—review & editing, H.L.; Project administration, H.L.; Funding acquisition, J.L. and P.L. All authors have read and agreed to the published version of the manuscript.

Funding: The work described in this paper is supported by the National Natural Science Foundation of China (Grant No. 62003375). The authors fully appreciate their financial supports.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

a	acceleration, $\mathrm{m/s^2}$
$\mathbf{A}, \mathbf{B}$	state-space model of the linearized equations of motion
H	Hamiltonian
$\mathbf{I}$	identity matrix
$J(\cdot)$	cost function
$\mathbf{L}$	constant vector
L^{-1}	inverse Laplace transform
LOS	light-of-sight
$Q(\cdot)$	reward signal
r	reward signal
s	state defined in Markov decision process
o	observation of the agent
$t, t_{\mathrm{go}}, t_{\mathrm{f}}$	time, time to go, and final time, respectively, s
u	guidance command, $\mathrm{m/s^2}$
V	velocity, m/s
$X-O-Y$	Cartesian reference frame
$\mathbf{x}$	state vector of the linearized equations of motion
y	lateral distance, m
Z	zero-effort-miss, m
α, β, σ	design parameters of the reward function
ϕ	flight path angle, rad
$\mathbf{\Phi}$	transition matrix
γ	discount factor
η	killing radius, m
λ	the angle between the corresponding light-of-sight and X-axis, rad
$\lambda(\cdot)$	Lagrange multiplier vector
$\mu(\cdot)$	policy function
ρ	relative distance between the adversaries, m
τ	time constant
ω, ξ	design parameters of the optimal guidance law (OGL)
I, T, D	interceptor, target, and defender, respectively
max	maximum
$*$	optimal solution

References

1. Ye, D.; Shi, M.; Sun, Z. Satellite proximate pursuit-evasion game with different thrust configurations. *Aerosp. Sci. Technol.* **2020**, *99*, 105715. [CrossRef]
2. Boyell, R.L. Defending a moving target against missile or torpedo attack. *IEEE Trans. Aerosp. Electron. Syst.* **1976**, *AES-12*, 522–526. [CrossRef]

3. Rusnak, I. Guidance laws in defense against missile attack. In Proceedings of the 2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 3–5 December 2008; pp. 090–094.
4. Rusnak, I. The lady, the bandits and the body guards—A two team dynamic game. *IFAC Proc. Vol.* **2005**, *38*, 441–446. [CrossRef]
5. Shalumov, V. Optimal cooperative guidance laws in a multiagent target–missile–defender engagement. *J. Guid. Control Dyn.* **2019**, *42*, 1993–2006. [CrossRef]
6. Weiss, M.; Shima, T.; Castaneda, D.; Rusnak, I. Combined and cooperative minimum-effort guidance algorithms in an active aircraft defense scenario. *J. Guid. Control Dyn.* **2017**, *40*, 1241–1254. [CrossRef]
7. Weiss, M.; Shima, T.; Castaneda, D.; Rusnak, I. Minimum effort intercept and evasion guidance algorithms for active aircraft defense. *J. Guid. Control Dyn.* **2016**, *39*, 2297–2311. [CrossRef]
8. Shima, T. Optimal cooperative pursuit and evasion strategies against a homing missile. *J. Guid. Control. Dyn.* **2011**, *34*, 414–425. [CrossRef]
9. Perelman, A.; Shima, T.; Rusnak, I. Cooperative differential games strategies for active aircraft protection from a homing missile. *J. Guid. Control Dyn.* **2011**, *34*, 761–773. [CrossRef]
10. Liang, H.; Wang, J.; Wang, Y.; Wang, L.; Liu, P. Optimal guidance against active defense ballistic missiles via differential game strategies. *Chin. J. Aeronaut.* **2020**, *33*, 978–989. [CrossRef]
11. Anderson, G.M. Comparison of optimal control and differential game intercept missile guidance laws. *J. Guid. Control* **1981**, *4*, 109–115. [CrossRef]
12. Dong, J.; Zhang, X.; Jia, X. Strategies of pursuit-evasion game based on improved potential field and differential game theory for mobile robots. In Proceedings of the 2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control, Harbin, China, 8–10 December 2012; pp. 1452–1456.
13. Li, Z.; Wu, J.; Wu, Y.; Zheng, Y.; Li, M.; Liang, H. Real-time Guidance Strategy for Active Defense Aircraft via Deep Reinforcement Learning. In Proceedings of the NAECON 2021-IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 16–19 August 2021; pp. 177–183.
14. Liang, H.; Li, Z.; Wu, J.; Zheng, Y.; Chu, H.; Wang, J. Optimal Guidance Laws for a Hypersonic Multiplayer Pursuit-Evasion Game Based on a Differential Game Strategy. *Aerospace* **2022**, *9*, 97. [CrossRef]
15. Liu, F.; Dong, X.; Li, Q.; Ren, Z. Cooperative differential games guidance laws for multiple attackers against an active defense target. *Chin. J. Aeronaut.* **2022**, *35*, 374–389. [CrossRef]
16. Weintraub, I.E.; Cobb, R.G.; Baker, W.; Pachter, M. Direct methods comparison for the active target defense scenario. In Proceedings of the AIAA Scitech 2020 Forum, Orlando, FL, USA, 6–10 January 2020; p. 0612.
17. Shalumov, V. Cooperative online guide-launch-guide policy in a target-missile-defender engagement using deep reinforcement learning. *Aerosp. Sci. Technol.* **2020**, *104*, 105996. [CrossRef]
18. Liang, H.; Wang, J.; Liu, J.; Liu, P. Guidance strategies for interceptor against active defense spacecraft in two-on-two engagement. *Aerosp. Sci. Technol.* **2020**, *96*, 105529. [CrossRef]
19. Salmon, J.L.; Willey, L.C.; Casbeer, D.; Garcia, E.; Moll, A.V. Single pursuer and two cooperative evaders in the border defense differential game. *J. Aerosp. Inf. Syst.* **2020**, *17*, 229–239. [CrossRef]
20. Harel, M.; Moshaiov, A.; Alkaher, D. Rationalizable strategies for the navigator–target–missile game. *J. Guid. Control Dyn.* **2020**, *43*, 1129–1142. [CrossRef]
21. Miljković, Z.; Mitić, M.; Lazarević, M.; Babić, B. Neural network reinforcement learning for visual control of robot manipulators. *Expert Syst. Appl.* **2013**, *40*, 1721–1736. [CrossRef]
22. Ye, D.; Chen, G.; Zhang, W.; Chen, S.; Yuan, B.; Liu, B.; Chen, J.; Liu, Z.; Qiu, F.; Yu, H. Towards playing full moba games with deep reinforcement learning. *arXiv* **2020**, arXiv:2011.12692.
23. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv* **2016**, arXiv:1610.03295.
24. Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J.J.; Gupta, A.; Fei-Fei, L.; Farhadi, A. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May 2017–3 June 2017; pp. 3357–3364.
25. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
26. Gaudeta, B.; Furfaroa, R.; Linares, R. Reinforcement meta-learning for angle-only intercept guidance of maneuvering targets. In Proceedings of the AIAA Scitech 2020 Forum AIAA 2020, Orlando, FL, USA, 6–10 January 2020; Volume 609.
27. Gaudet, B.; Linares, R.; Furfaro, R. Adaptive guidance and integrated navigation with reinforcement meta-learning. *Acta Astronaut.* **2020**, *169*, 180–190. [CrossRef]
28. Lau, M.; Steffens, M.J.; Mavris, D.N. Closed-loop control in active target defense using machine learning. In Proceedings of the AIAA Scitech 2019 Forum, San Diego, CA, USA, 7–11 January 2019; p. 0143.
29. Zhang, G.; Chang, T.; Wang, W.; Zhang, W. Hybrid threshold event-triggered control for sail-assisted USV via the nonlinear modified LVS guidance. *Ocean Eng.* **2023**, *276*, 114160. [CrossRef]
30. Li, J.; Zhang, G.; Shan, Q.; Zhang, W. A novel cooperative design for USV-UAV systems: 3D mapping guidance and adaptive fuzzy control. *IEEE Trans. Control Netw. Syst.* **2022**, *10*, 564–574. [CrossRef]

31. Ainsworth, M.; Shin, Y. Plateau phenomenon in gradient descent training of RELU networks: Explanation, quantification, and avoidance. *SIAM J. Sci. Comput.* **2021**, *43*, A3438–A3468. [CrossRef]
32. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In *PMLR, Proceedings of Machine Learning Research, Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; PMLR: New York, NY, USA, 2018; Volume 80, pp. 1587–1596.
33. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
34. Babaeizadeh, M.; Frosio, I.; Tyree, S.; Clemons, J.; Kautz, J. Reinforcement learning through asynchronous advantage actor-critic on a gpu. *arXiv* **2016**, arXiv:1611.06256.
35. Casas, N. Deep deterministic policy gradient for urban traffic light control. *arXiv* **2017**, arXiv:1703.09035.
36. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic Policy Gradient Algorithms. In *PMLR, Proceedings of Machine Learning Research, Proceedings of the 31st International Conference on Machine Learning, Beijing China, 21–26 June 2014*; PMLR: New York, NY, USA, 2014; Volume 32, pp. 387–395.
37. Fan, J.; Wang, Z.; Xie, Y.; Yang, Z. A Theoretical Analysis of Deep Q-Learning. In *PMLR, Proceedings of Machine Learning Research, Proceedings of the 2nd Conference on Learning for Dynamics and Control, Online, 10–11 June 2020*; PMLR: New York, NY, USA, 2020; Volume 120, pp. 486–489.
38. Hasselt, H. Double Q-learning. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: New York, NY, USA, 2010; Volume 23.
39. Gullapalli, V.; Barto, A.G. Shaping as a method for accelerating reinforcement learning. In Proceedings of the 1992 IEEE International Symposium on Intelligent Control, Glasgow, UK, 11–13 August 1992; pp. 554–559.
40. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 41–48.
41. Krueger, K.A.; Dayan, P. Flexible shaping: How learning in small steps helps. *Cognition* **2009**, *110*, 380–394. [CrossRef]
42. Ng, A.Y.; Harada, D.; Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. *LCML* **1999**, *99*, 278–287.
43. Randløv, J.; Alstrøm, P. Learning to Drive a Bicycle Using Reinforcement Learning and Shaping. *ICML* **1998**, *98*, 463–471.
44. Wiewiora, E. Potential-based shaping and Q-value initialization are equivalent. *J. Artif. Intell. Res.* **2003**, *19*, 205–208. [CrossRef]
45. Qi, N.; Sun, Q.; Zhao, J. Evasion and pursuit guidance law against defended target. *Chin. J. Aeronaut.* **2017**, *30*, 1958–1973. [CrossRef]
46. Ho, Y.; Bryson, A.; Baron, S. Differential games and optimal pursuit-evasion strategies. *IEEE Trans. Autom. Control* **1965**, *10*, 385–389. [CrossRef]
47. Shinar, J.; Steinberg, D. Analysis of Optimal Evasive Maneuvers Based on a Linearized Two-Dimensional Kinematic Model. *J. Aircr.* **1977**, *14*, 795–802. [CrossRef]