

INTEGRATED CONTROL STRATEGY BASED ON PHYSICS-INFORMED MODEL-BASED REINFORCEMENT LEARNING FOR AIRBORNE HIGH-ENERGY SYSTEMS

Shilong Bai[1,2], Yang Li[1,2], Linlin Jiang[1,2], Zhuo Li[1,2], Ke He[1,2] & Zongxia Jiao[1,2]

[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China
[2]National Key Laboratory of Aircraft Integrated Flight Control, Beihang University, Beijing 100191, China

Abstract

Airborne high-energy systems (HESs) provide rapid energy delivery and deep-magazine potential for future air combat, but their tactical use is strongly constrained by airborne thermal-management capacity. Existing air-combat strategy studies rarely couple laser engagement decisions with the dynamic thermal state of the weapon-platform system, which limits the credibility of operational effectiveness evaluation. This paper develops an integrated HES air-combat control strategy based on physics-informed model-based reinforcement learning. A unified energy-information simulation environment is first established by coupling aircraft dynamics, missile engagement, radar/fire-control logic, laser damage, and HES thermal-management models. On this basis, a hierarchical reinforcement-learning framework is designed: the high-level Actor-Critic policy generates maneuver and HES activation decisions, while a low-level PID controller executes stable flight control and the fire-control module enforces thermal engagement constraints. The HES heat-sink residual capacity is embedded into both the state representation and reward design, enabling the agent to learn laser use under thermal safety margins. In 100 randomized simulations, the HES integrated countermeasure policy achieves a 64% win rate against a pursuit-attack policy and a 55% win rate with an 18% failure rate against a missile integrated countermeasure policy. The results indicate that the proposed framework can support closed-loop evaluation of flight-fire integrated HES strategies under dynamic thermal constraints.

Keywords: Airborne high-energy systems; Thermal constraint; Physics-informed learning; Integrated control strategy; Air combat simulation

1. Introduction

With the evolution of modern air combat towards high-dynamic and intense confrontation environments, airborne high-energy systems (HESs), using technical features such as light-speed engagement, sustained firing capacity, and precision damage, are emerging as promising systems that may reshape traditional air combat paradigms [1, 2]. Unlike conventional kinetic weapons, HESs achieve target destruction through high-energy laser beams, offering multi-level strike capabilities ranging from non-lethal interference to hard-kill effects, while maintaining precision engagement of vulnerable components on enemy equipment even in strong electromagnetic interference environments. Additionally, HESs primarily rely on electrical power, reducing per-use costs while eliminating traditional ammunition loads, thereby enhancing operational platforms' endurance and payload efficiency. HES combat differs from traditional weapons in strategy, damage mechanisms, and operational effectiveness, necessitating specialized experimental evaluation methods. For instance, the instantaneous thermal load generated during high-power laser operation (e.g., 80 kW waste heat at 200 kW laser activation) conflicts with limited aircraft cooling capacity [3, 4]. Experimental data indicate that

https://doi.org/10.71945/icas2026_0621

when laser diode junction temperatures exceed 35°C, weapon efficiency decreases by over 30%, directly reducing lethal range [5, 6]. This energy-thermal coupling effect severely constrains sustained combat capability, rendering conventional tactical strategies inadequate under thermal constraints.

Research on HES air-combat strategy and effectiveness evaluation forms a critical component of autonomous decision-making for high-energy weapon systems. Previous studies have examined tactical parameters and laser design variables for HEL-equipped aircraft through simulation [7]. These studies show that altitude, speed, pointing accuracy, and laser power influence damage effectiveness, but they often assume simplified target motion and do not couple tactical decisions with onboard thermal-management constraints.

Reinforcement learning is well suited to air-combat decision-making because it learns through environmental interaction and reward feedback. Deep reinforcement learning has been applied to aircraft control, hierarchical air-combat decision-making, reward-structure design, and multi-agent maneuvering [8, 9, 10, 11]. Existing methods improve maneuver decision-making, but most focus on aircraft control or generic policy selection and do not explicitly integrate HES thermal state, laser fire-control feasibility, and tactical decision-making in a closed-loop environment.

This paper addresses HES combat control strategy challenges by developing an integrated energy-information simulation environment that bridges operational effectiveness evaluation and strategy validation. In this study, the term model-based refers to the use of physics-based aircraft, missile, laser-damage, and thermal-management models to construct the decision environment and state feedback, rather than to a learned transition model in the narrow sense of model-based reinforcement learning. The proposed strategy is therefore physics-informed: the learned policy is optimized in a model-supported combat environment while explicitly observing and responding to the HES thermal state.

This work makes three main contributions. First, an integrated simulation environment is established to couple flight dynamics, missile engagement, radar/fire-control logic, laser damage, and HES thermal management, enabling closed-loop evaluation of combat strategy under dynamic thermal constraints. Second, a hierarchical flight-fire control strategy is designed in which the high-level policy generates discrete maneuver and fire-control decisions, the low-level PID controller converts maneuver decisions into stable aircraft control actions, and the fire-control module enforces weapon engagement conditions. Third, the HES heat-sink residual capacity is incorporated into both the policy state and the HES activation reward, allowing the agent to learn anti-missile laser use under thermal safety constraints. Compared with previous hierarchical reinforcement-learning studies that mainly focus on maneuver control or generic policy selection, this work emphasizes the coupling between HES-specific thermal constraints and air-combat fire-control decisions.

2. Air combat environment model

HESs are advanced weapon systems that employ concentrated electromagnetic energy as their primary means of engagement, delivering precise strikes or damage through high-power energy output. Operating in complex air combat environments, HES mission effectiveness depends on multiple factors. To validate air combat strategy design, an HES air-combat simulation environment is constructed, comprising ego aircraft, enemy aircraft (or missile), and environmental elements. This modular simulation environment replicates real-world battlefield physics and includes submodels for HES airborne platforms and operational workflows, such as aircraft model, thermal model, missile model, radar detection model, and damage model.

2.1 Dynamic thermal model

The operational process of the HES considered in this work consists of energy generation, energy transmission, and energy output. The core advantages of HESs, such as instantaneous lethality and sustained combat capability, are accompanied by challenges posed by instantaneous thermal loads under high-power operation. Fiber laser serves as the primary component for energy conversion, involving electrical-to-optical and optical-to-optical transformation. Key components of fiber laser include laser diode, pump combiner, gain fiber, and power combiner. Due to electrical-optical conversion losses and transmission inefficiencies, fiber laser generates nearly 90% of the total heat within the high-energy laser system, making it the dominant thermal source.

The unique constraints of airborne environments (limited space, complex aerodynamic conditions, and strict weight limitations) mean that ineffective thermal management of HES-generated heat during prolonged operation directly impacts equipment performance stability and flight safety. The airborne thermal management system, a critical subsystem in modern aircraft, manages heat from airborne devices via efficient cooling circuits to maintain safe operating temperatures [12, 13, 14]. After integration, HES becomes the largest thermal sink consumer of the aircraft's thermal management system, particularly as next-generation aircraft face megawatt-level heat loads from high-energy systems [15]. Figure 1 shows the HES thermal-management sequence considered in this work, where waste heat is transferred from HES to thermal storage, then to the heat exchanger and aircraft heat sink [16, 17].

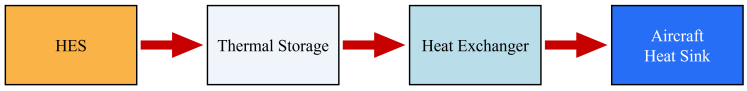


Figure 1 – Thermal management system for HES.

2.1.1 Heat generation model

High-energy laser weapon typically employs fiber laser architecture, in which double-clad fiber laser utilizes high-power laser diode arrays as pump source to achieve efficient high-power output. The emitted laser forms a near-perfect Gaussian beam, simulated using Gaussian curve approximations [18].

During laser emission, the laser diode generates heat, causing temperature variations in core components that directly impact performance. Rising temperatures degrade the output optical power of laser diode and induce wavelength drift in the emitted spectrum.

The output optical power P_{out} of the laser diode decreases as its junction temperature rises above a specified reference temperature T_{ref}. Eqs. (1)–(2) describe how the temperature-dependent threshold current I_{th} and slope efficiency η govern the laser diode's output characteristics [19].

$$P_{\text{out}} = \eta \times (I - I_{\text{th}}) \tag{1}$$

$$\begin{cases} I_{\text{th}}(T_j) = I_{\text{ref}} \exp\left(\dfrac{T_j - T_{\text{ref}}}{T_0}\right) \\ \eta(T_j) = \eta_{\text{ref}} \exp\left(-\dfrac{T_j - T_{\text{ref}}}{T_1}\right) \end{cases} \tag{2}$$

where I_{ref} is the threshold current of the laser diode at a reference temperature T_{ref}, and η_{ref} represents the slope efficiency under the same reference conditions. T_0 (threshold characteristic temperature) and T_1 (slope characteristic temperature) are experimentally determined parameters specific to individual laser diode. When maintaining a constant current I, the concurrent decrease in η and increase in I_{th} induced by temperature elevation collectively result in reduced P_{out}.

Temperature-induced spectral drift and combiner loss further reduce the optical power delivered to the target. Because these optical details are not the main focus of the control strategy, the simulation aggregates laser diode efficiency, gain-fiber absorption, and combiner efficiency into the HES heat-generation calculation. All laser energy not ultimately delivered to the target for strike or damage is considered wasted and converted into thermal energy.

2.1.2 Thermal management model

The thermal management system comprises three key stages, including the thermal interface between HES and thermal storage unit, the thermal storage unit, and thermal exchange between the thermal storage unit and aircraft platform thermal sink. Equivalent thermal capacitance C and equivalent thermal resistance R parameters are established for each heat transfer stage, enabling quantitative analysis of component temperatures and heat flux dynamics through governing thermal equations.

$$\Delta T = \frac{q_{in} - q_{out}}{C} \cdot \Delta t \tag{3}$$

$$q = \frac{\Delta T}{R} \tag{4}$$

The thermal-management subsystem is represented as a lumped resistance-capacitance network rather than a detailed component-level heat-transfer model. This abstraction is sufficient for the present conference paper because the control policy only needs the dynamic heat-sink margin and its influence on feasible HES activation, not the detailed temperature field inside each cooling component.

The heat-transfer path follows three stages: heat generated by the laser diode is first removed through a microchannel cold plate, then buffered by a phase-change thermal storage unit, and finally rejected to the aircraft heat sink through an evaporative heat-exchange loop [20, 21, 22]. Rectangular microchannels are used because of their compact structure and favorable temperature uniformity in high-power laser cooling applications. Phase-change material is introduced as an intermediate thermal buffer to absorb the intense transient waste heat produced during HES activation [23, 24].

In the simulation, the above stages are aggregated into equivalent thermal resistances and thermal capacitances. The resulting heat-sink residual capacity is updated at each decision step and passed to the reinforcement-learning agent as η_{cool}. When HES activation consumes thermal capacity, η_{cool} decreases; during the cooling phase, heat rejection to the aircraft heat sink allows it to recover. This variable therefore provides the control policy with a compact representation of the current thermal safety margin [25].

2.2 Aircraft model

The F-16 fighter aircraft is chosen as the primary model for air combat simulation due to its balanced performance, extensive operational deployment, and mature technical framework. The aircraft model integrates two core components: flight dynamics/kinematics model and flight control system.

For the flight dynamics/kinematics model, JSBSim (an open-source flight dynamics simulation framework) is employed to meet high-fidelity simulation requirements. It has also been used in RL-based air-combat environments [26]. JSBSim is widely recognized in aerospace research for its precision in simulating 6DOF motion equations, enabling realistic replication of aircraft behavior across diverse flight conditions. It supports customizable models, accurate atmospheric environment simulation, and seamless integration with tools like FlightGear and Tacview, ensuring both authenticity and interoperability in simulation analysis.

The F-16 control system adopts a dual-loop structure. The outer loop begins with the guidance module generating target parameters, including velocity V_{target}, heading ψ_{target}, and altitude h_{target}. These targets feed into PID-based outer-loop controllers, which compute command errors to derive inner-loop control signals, including throttle setting, roll, yaw, and pitch commands. The inner loop focuses on rapid attitude stabilization through direct actuation of control surfaces, ensuring precise attitude adjustment and dynamic stability.

2.3 Missile model

The missile model employs the AIM-9L as the representative short-range air-to-air missile. For the conference-level simulation description, the missile is modeled using point-mass dynamics with powered and unpowered flight phases, while its terminal guidance follows proportional navigation in the line-of-sight azimuth and elevation channels. This model provides missile position, velocity, time-to-intercept tendency, and hit/miss outcomes for the engagement simulation.

In the air-combat environment, the missile model is not intended to reproduce every subsystem of a tactical missile. Instead, it provides a physically consistent threat model for evaluating whether the aircraft can evade incoming missiles or combine maneuvering, missile launch, and HES interception under dynamic engagement conditions.

2.4 Radar detection model

The radar model provides target detection and lock information for both missile launch and HES engagement. During combat operations, the radar first operates in search mode. Once the target satisfies detection and tracking conditions, it switches to attack mode and provides the fire-control module with relative range, angular information, and lock status.

In the simulation, radar performance is represented by operational gates rather than by a detailed radar-equation calculation. These gates include maximum detection range, azimuth and elevation coverage, lock duration, and engagement-mode availability. The resulting radar state determines whether missile launch or HES activation commands can be converted into executable fire-control decisions.

2.5 Damage model

During atmospheric propagation, the laser beam is degraded by multiple factors, including platform jitter, atmospheric diffraction, scattering, turbulence, and thermal blooming. These effects collectively enlarge the beam spot on the target, reducing energy density.

The ideal energy density is used only as a reference. In practice, platform vibrations, mechanical friction, and sensor errors cause the laser spot to distribute uniformly around the aimpoint, linearly enlarging the spot radius to a_t. Turbulence and thermal blooming effects are considered negligible here, with primary losses attributed to atmospheric absorption, scattering, and diffraction. Diffraction further enlarges the beam radius by a_d.

The actual on-target energy density I_{actual} incorporates atmospheric transmittance τ_a, optical system efficiency τ_0, target reflectivity r, and laser incidence angle α.

$$I_{\text{actual}} = \frac{P_s \tau_a \tau_0 (1-r)\cos\alpha}{\pi(a_t^2 + a_d^2)} \tag{5}$$

Laser weapons prioritize striking components with low energy thresholds and minimal shielding. The damage effect is quantified by the dwell time Δt required to penetrate a material layer.

$$\Delta t = \frac{\rho_T h_T c_T (T_T - T_0) + \rho_T h_T H_T}{I_{\text{actual}}} \tag{6}$$

where T_0 is the initial temperature, T_T the melting temperature, ρ_T the material density, h_T the thickness, c_T the specific heat capacity, and H_T the latent heat of fusion.

3. Integrated control strategy based on hierarchical reinforcement learning

Hierarchical Reinforcement Learning (HRL) is a methodology designed to address high-dimensional complex tasks by incorporating a layered structural framework. Its core principle lies in task decomposition, where a global objective is partitioned into a hierarchy of subtasks, thereby reducing the dimensionality of the decision space and enhancing learning efficiency.

3.1 Strategy training framework

Figure 2 shows the training framework for red-blue force tactical engagement in an air combat simulation environment. The red aircraft employs an HRL-based intelligent decision-making framework with three explicitly separated levels. The high-level policy is implemented by an Actor-Critic network and receives normalized combat states, including relative engagement geometry, missile threat information, and HES thermal status. It outputs discrete maneuver intentions, namely target altitude, heading, and velocity increments, together with the learned HES activation command. The low-level PID controller then tracks the maneuver intentions and converts them into continuous aircraft control commands. The fire-control decision module evaluates weapon activation using the high-level command together with physical engagement gates, such as attack angle, target distance, lock duration, firing interval, missile inventory, and HES heat-sink residual capacity.

In this framework, weapon command denotes the discrete intent generated at the decision layer, whereas fire-control decision denotes the executable engagement result after applying physical and operational constraints. In the HES integrated policy used in this work, missile firing is enabled by

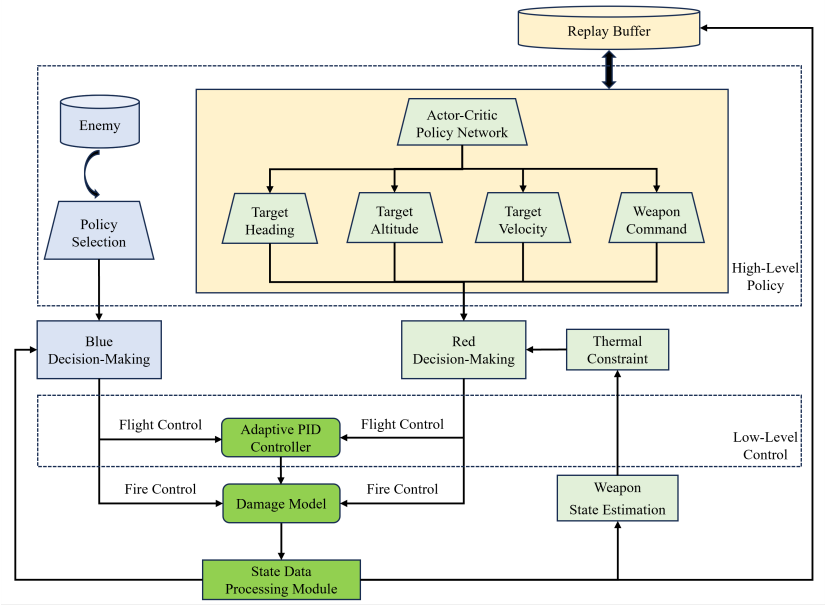


Figure 2 – Strategy training framework for HRL-based tactical engagement.

rule-based engagement gates, while HES activation is learned by the high-level policy and further constrained by the thermal and geometric gates. The blue aircraft operates on a rule-driven policy framework, generating countermeasures by parsing the blue decision data. Its adaptive PID controller ensures flight attitude stability, while the damage model quantifies combat effectiveness after fire-control execution. Both systems feed operational feedback into the state data processing module, forming a closed-loop framework for continuous battlefield situational awareness and iterative strategy optimization.

3.2 Air combat engagement model

The air combat engagement model provides input and evaluation criteria for agent decision-making by quantifying situational superiority between opposing forces, as shown in Figure 3.

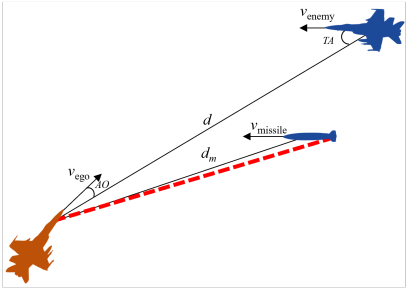


Figure 3 – The situational superiority of air combat engagement model.

In this model, the positions, velocities, attitudes, and angular rates of the ego aircraft, enemy aircraft, and missiles are tracked in three-dimensional space. The engagement state is summarized through relative distance, relative velocity, aspect angle, tail angle, lateral position flags, missile threat geometry, and platform health state.

Operational survivability is represented by a scalar health state (0-100), where a value less than 0 indicates platform elimination. This variable is used for damage accumulation and terminal-state judgment; missile hits are treated as direct destructive events in the combat outcome evaluation. The resulting engagement quantities are then used for state construction, reward shaping, fire-control gating, and win/failure/draw judgment.

3.3 State space and action space

The state space comprises three categories: ego aircraft status, relative situational data, and HES status. The selected state variables are summarized in Table 1.

Table 1 – State space

Symbol	Scale/Unit	Description	Symbol	Scale/Unit	Description
h	5 km	Ego altitude	d	10 km	Relative distance to enemy
$\sin\phi$	–	Sine of roll angle	s_f	–	Enemy lateral position flag
$\sin\theta$	–	Sine of pitch angle	Δv_{mx}	m/s	Relative X-axis velocity difference (missile)
$\cos\phi$	–	Cosine of roll angle	Δh_m	1 km	Relative altitude difference (missile)
v_x	m/s	Ego X-axis velocity	AO_m	rad	Ego aspect angle to missile
v_y	m/s	Ego Y-axis velocity	TA_m	rad	Ego bearing angle to missile
v_z	m/s	Ego Z-axis velocity	d_m	10 km	Distance to missile
v_c	m/s	Ego total velocity	s_m	–	Lateral position flag (missile)
AO	rad	Ego aspect angle	η_{cool}	–	Heat sink residual capacity
TA	rad	Ego bearing angle			

Aspect angle and tail angle are computed from the relative line-of-sight geometry between the platform and the target, while the lateral position marker indicates the target's bearing relative to the platform's motion direction: 1 denotes the target being on the left side, -1 on the right side, and 0 aligned with the motion direction. It is noteworthy that the selected state space involves multiple dimensions with varying value magnitudes. To ensure policy precision, this paper normalizes all state space values to the range of -10 to 10.

The missile-related policy uses the aircraft state, relative engagement state, and missile-threat state listed in Table 1. For aircraft carrying HES, the state space is augmented by the additional thermal-state variable η_{cool}, which denotes the normalized heat-sink residual capacity produced by the HES thermal-management model. This variable closes the loop between thermal dynamics and policy learning: when HES activation increases the diode temperature and consumes heat-sink capacity, η_{cool} decreases in the next observation, thereby informing subsequent maneuver and fire-control decisions.

As mentioned in the training framework, the agent's decision commands comprise two categories: flight control commands and fire control commands. Flight control commands can be either discrete or continuous. Although aircraft flight control fundamentally constitutes a continuous action problem, learning control strategies in three-dimensional continuous action spaces presents significant challenges. During initial training phases, the stochastic nature of policies makes it difficult for the agent to generate sufficient exploratory behavior trajectories, while control command oscillations remain hard to eliminate. Therefore, the proposed hierarchical approach in this paper designs the flight control command space as discrete variables, as detailed in Table 2.

The discrete flight control action space comprises aircraft altitude change, heading angle change, and velocity change. In implementation, the velocity command is represented by normalized policy outputs $(-0.2, 0, 0.2)$ and then denormalized to the physical velocity increments shown in Table 2. Compared to flight control action spaces constructed with predefined maneuver sets, the discrete action space proposed in this paper demonstrates enhanced action diversity. These command outputs maintain stable flight when executed by the underlying adaptive PID controller.

Fire control commands are discrete variables governing weapon deployment. Eq. (7) defines the fire-control command group, which contains the missile fire command a_{missile} and the HES fire command

Table 2 – Discrete flight control commands

Action name	Range	Unit	Description
Δh	(-0.2, 0, 0.2)	km	Altitude variation
$\Delta\psi$	(-30°, -15°, 0°, 15°, 30°)	°	Heading angle variation
Δv	(-68, 0, 68)	m/s	Velocity command increment

a_{HES}. For each command, value 1 indicates authorized activation and value 0 indicates prohibited deployment.

$$a_{\text{shoot}} = (a_{\text{missile}}, a_{\text{HES}}) \tag{7}$$

In implementation, this fire-control vector represents the logical command set, while the final executable firing decision is produced after engagement gating. A missile launch is allowed only when the target is within the prescribed attack angle and distance, lock duration is satisfied, the launch interval has elapsed, and remaining missile inventory is positive. HES activation additionally requires an incoming missile target, sufficient lock condition, range satisfaction, firing-interval satisfaction, and $\eta_{\text{cool}} > 0.2$. Thus, the high-level policy proposes fire-control intent, and the environment converts it into physically admissible weapon execution.

The complete action space for air combat scenarios therefore combines the discrete flight-control commands with the weapon fire-control commands.

3.4 Reward function

For air combat scenarios, this study designs five reward functions, including altitude reward, missile evasion reward, pursuit reward, HES countermeasure reward, and event reward. These reward terms are not executed as separate objectives. Instead, they are evaluated at each decision step and combined as a weighted scalar reward for PPO optimization. The altitude and event rewards impose safety and terminal-event guidance, the missile evasion and pursuit rewards provide dense situational shaping, and the HES countermeasure reward introduces laser-use preference under geometric and thermal constraints.

(1) Altitude reward

The altitude reward penalizes aircraft flying below safe altitudes to prevent intentional crashes. The safety altitude is set at 4 km, with a danger threshold at 3.5 km.

When the aircraft descends below the safe altitude, a penalty is imposed according to both the current altitude and the vertical velocity, discouraging continued descent near the lower boundary. When the altitude falls below the danger threshold or the aircraft crashes, a stronger penalty is applied through the safety and terminal-event rewards. The resulting altitude reward term is denoted as R_h. This design provides simple flight-safety guidance without introducing an additional optimization objective.

(2) Missile Evasion Reward

The missile evasion reward function incentivizes aircraft to evade incoming missiles through dual considerations of guidance effectiveness and vertical velocity differential. Missile guidance effectiveness is quantified by the cosine of the angle between the missile's velocity vector and the line of sight vector.

$$r_{LOS} = -\operatorname{clip}\left(\frac{v_m \cdot LOS_{\text{unit}}}{\|v_m\|}, -1, 1\right) \tag{8}$$

where clip denotes the clipping function, $\vec{v}_m$ represents the velocity vector of the incoming missile, and LOS_{unit} is the unit line-of-sight vector from the missile to the aircraft, computed from their relative position. A smaller angle indicates superior missile guidance performance and consequently higher penalties due to reduced lateral evasion capability.

The vertical velocity differential term r_{vz} is obtained by clipping the difference between the aircraft and missile vertical velocities, v_{az} and v_{mz}, to the range of $[-1, 1]$. Smaller aircraft vertical velocity relative

to the missile results in increased penalties due to diminished vertical escape capacity. Eq. (9) gives the composite missile evasion reward.

$$R_{\text{evade}} = 0.6 \cdot r_{LOS} + 0.4 \cdot r_{vz} \tag{9}$$

(3) Pursuit reward

The pursuit reward mechanism incentivizes proper heading alignment toward enemy aircraft while penalizing vulnerable pursuit scenarios, primarily encouraging the maintenance of advantageous combat positions and preventing passive engagement. This reward component integrates both aspect angle dominance and distance-related penalties.

$$R_{\text{pursue}} = \cos(AO) - \cos(TA) - d \tag{10}$$

where AO quantifies the angular offset between the velocity vector and the line-of-sight of ego aircraft to the enemy aircraft, TA inversely measures the enemy's angular advantage relative to the ego, and d denotes the relative distance between aircraft. Tactical superiority emerges when the ego aircraft achieves tail-chase geometry with proper heading alignment toward the enemy aircraft while maintaining a shorter engagement distance. Conversely, elevated TA values reflect enemy's angular dominance, signaling defensive vulnerability as enemy gains favorable firing solutions.

(4) HES countermeasure reward

The HES countermeasure reward function incentivizes aircraft to activate the laser weapon under optimal operational conditions by synthesizing distance, aspect angle, and thermal constraint. Eq. (11) defines this reward.

$$R_{\text{HES}} = a_{\text{HES}} \cdot \text{clip}\left(\frac{|\cos(AO_m)|}{d_m} + (\eta_{\text{cool}} - 0.2), -10, 10\right) \tag{11}$$

where a_{HES} denotes the executable HES fire-control result, AO_m represents the aspect angle between the aircraft and incoming missile, d_m specifies their relative distance, and η_{cool} quantifies the HES system's heat-sink residual capacity. The reward escalates as the aircraft-missile proximity increases and their aspect angle diminishes, reflecting dual implications: such conditions simultaneously enhance the aircraft's capability for laser-based missile interception and amplify collision risks. The thermal term introduces an explicit preference for firing when sufficient cooling margin is available. When η_{cool} approaches the lower bound of 0.2, the term $(\eta_{\text{cool}} - 0.2)$ becomes close to zero or negative, reducing the reward for HES activation. In addition, the execution gate prohibits HES activation when $\eta_{\text{cool}} \leq 0.2$. Therefore, low heat-sink residual capacity suppresses HES use through both reward shaping and hard fire-control constraints.

(5) Event reward

The event reward mechanism imposes substantial rewards/punishments for critical combat occurrences to amplify long-term strategic objectives within the reinforcement learning framework. This reward component primarily depends on the operational status of ego and enemy missiles.

$$R_{\text{event}} = \begin{cases} +100, & \text{Enemy missile miss} \\ -200, & \text{Enemy missile hit} \\ -100, & \text{Ego missile miss} \\ +200, & \text{Ego missile hit} \end{cases} \tag{12}$$

Specifically, +100 reward is granted upon successful evasion of an incoming enemy missile, whereas -200 penalty is applied if the aircraft suffers a missile hit. Conversely, ego missile performance is evaluated through a symmetric structure: -100 penalty is incurred for failed missile engagements, while +200 reward is awarded upon successful target destruction. If multiple events occur within the same decision step, their corresponding event rewards are accumulated.

Eq. (13) gives the overall reward used by the HES integrated countermeasure policy.

$$R_{\text{total}} = w_h R_h + w_{\text{evade}} R_{\text{evade}} + w_{\text{pursue}} R_{\text{pursue}} + w_{\text{HES}} R_{\text{HES}} + w_{\text{event}} R_{\text{event}}. \tag{13}$$

In the HES integrated scenario, the implementation uses $w_h = 1$, $w_{\text{evade}} = 1$, $w_{\text{pursue}} = 5$, $w_{\text{HES}} = 10$, and $w_{\text{event}} = 1$. When reward terms conflict, flight safety and weapon feasibility are first enforced by hard environment constraints, such as altitude termination, missile engagement gates, and the HES thermal threshold $\eta_{\text{cool}} > 0.2$; within the feasible action set, the weighted reward guides the policy toward tactically favorable maneuvering and timely HES activation.

3.5 Actor-Critic-based policy network

The Actor-Critic (AC) method constitutes a reinforcement learning algorithm that integrates policy gradients with value functions [27]. Its fundamental principle lies in simultaneously learning both policy and value functions. Within this framework, the Actor represents the policy network responsible for action determination, while the Critic embodies the value network estimating state values. The AC method employs value feedback from the Critic to guide the Actor's policy optimization, thereby enhancing policy performance.

The designed policy network is based on the Actor-Critic method, as shown in Figure 4. The Actor and Critic share a common feature extraction layer comprising two fully connected layers and a Gated Recurrent Unit. The initial FC layer performs preliminary state feature extraction, the recurrent module captures temporal features, and the subsequent FC layer conducts advanced feature processing. The Actor maps extracted features to multi-dimensional action probabilities, whereas the Critic produces scalar value estimates through linear projection.

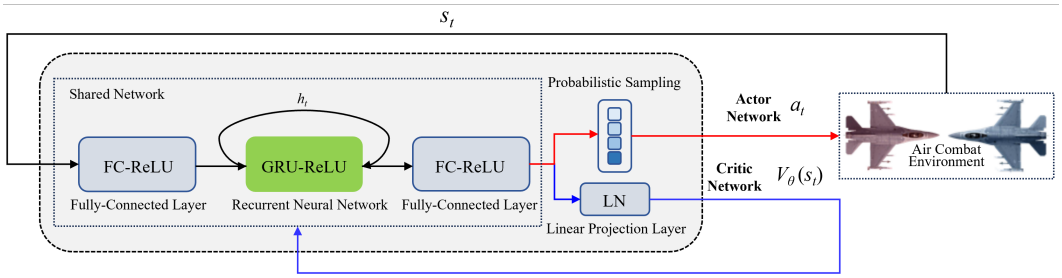


Figure 4 – Policy network framework based on Actor-Critic method.

Given the temporal dependencies inherent in air combat decision sequences, the intermediate GRU module captures temporal state characteristics through hidden state transmission between consecutive timesteps [28]. Compared with Long Short-Term Memory network, GRU is structurally simpler and more suitable for compact recurrent policy implementation in the present simulation.

3.6 Proximal policy optimization algorithm

The policy network is optimized using Proximal Policy Optimization (PPO), which improves training stability by limiting excessive policy updates [29]. In this work, PPO updates the high-level policy that generates maneuver intentions and HES activation commands, while the low-level controller and fire-control gates ensure physically feasible execution.

Let θ denote the Actor parameters and ϕ denote the Critic parameters. The Actor is optimized using the clipped surrogate objective

$$J^{\text{clip}}(\theta) = \mathbb{E}_t \left[\min \left(\rho_t(\theta) A_t, \text{clip}\left(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon\right) A_t \right) \right], \tag{14}$$

where A_t is the advantage estimate, ε is the clipping range, and $\rho_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ is the importance sampling ratio between the updated and old policies. The clipping operation constrains the policy update magnitude.

The Critic estimates state values for advantage calculation, and an entropy bonus is included to encourage exploration. The overall PPO training procedure is summarized in Algorithm 1.

Table 3 summarizes the main PPO training settings used in the simulations.

Algorithm 1 PPO in Actor-Critic Framework

Initialize Actor π_θ, Critic V_ϕ, replay buffer, and opponent pool
for each training iteration do
  Roll out $\pi_{\theta_{\text{old}}}$ in parallel environments and collect trajectories
  Estimate advantages and target returns from rewards and value predictions
  Split trajectories into recurrent mini-batches
  Update the Actor using the clipped PPO objective in Eq. (14)
  Update the Critic by minimizing value prediction error and apply gradient clipping
  Set $\theta_{\text{old}} \leftarrow \theta$ and update training records when required
end for

Table 3 – Training hyperparameters

Parameter	Value	Parameter	Value
Mini-batches per update	5	Replay buffer size	3000
Total environment steps	1×10^{8}	Rollout threads	32
Learning rate	3×10^{-4}	Discount factor γ	0.99
Data chunk length	8	PPO epochs per update	4
PPO clipping parameter	0.2	Gradient-norm clipping	2.0
Entropy coefficient	1.0		

4. Results and discussion

4.1 Trained air combat policy

Building upon the strategy framework proposed in Section 3, different air-combat policies with distinct operational objectives can be designed to simulate realistic air-combat scenarios. This study considers missile-evasion, pursuit-attack, missile integrated countermeasure, and HES integrated countermeasure policies.

The missile evasion policy aims to improve aircraft survivability by maneuvering against incoming missiles. Through evasive maneuvers, the enemy missile may detonate prematurely, deviate from the target, or consume the enemy missile inventory. In this setting, the ego aircraft does not launch missiles, and the action space only includes flight-control commands. The reward function uses the altitude reward, missile evasion reward, and event reward.

The pursuit-attack policy aims to keep the aircraft nose directed toward the enemy, close the engagement distance, and launch missiles to seize tactical initiative. The agent is equipped with one missile, and the action space includes both flight-control commands and missile fire-control commands. In the hierarchical framework, the interface between the high-level policy and the low-level PID controller is defined by heading, altitude, and velocity errors, which can be obtained directly from the relative engagement state. Missile fire-control execution follows predefined launch rules.

The missile integrated countermeasure policy requires the aircraft to make decisions according to the current engagement state. When the incoming missile threat is high, the aircraft performs evasive maneuvers; when an attack opportunity appears, it closes the distance, occupies an advantageous position, and launches a missile. The action space is consistent with the pursuit-attack policy, and the reward function integrates the altitude reward, missile evasion reward, pursuit reward, and event reward.

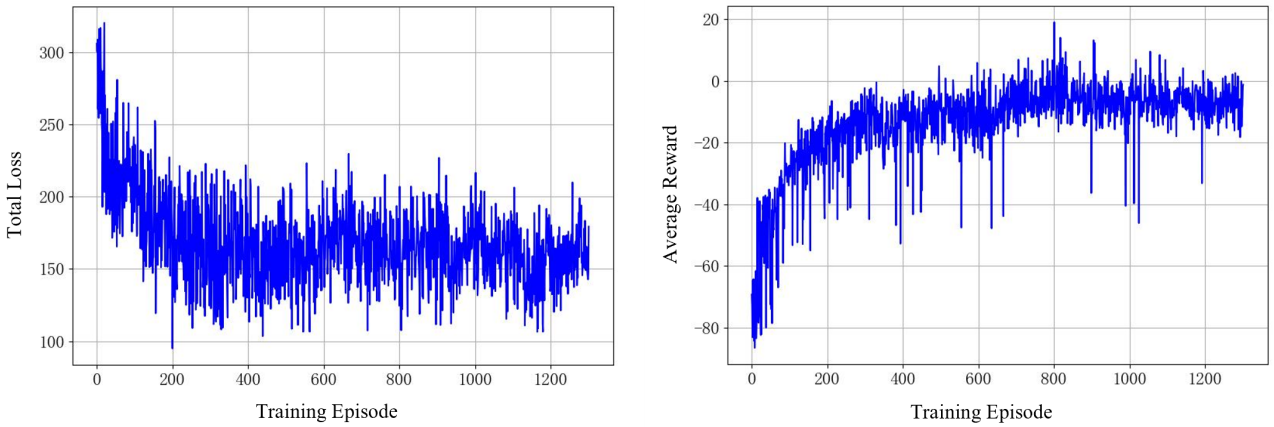
The HES integrated countermeasure policy requires the aircraft to make both maneuvering and fire-control decisions. When the incoming missile threat is high, the aircraft can evade by maneuvering or activate the HES to intercept the missile. It can also seize favorable engagement positions and launch

missiles against the enemy aircraft. This policy incorporates HES thermal constraints into the optimization objective and supports autonomous attack decisions under the thermal safety boundary. Its action space includes flight-control commands, missile fire-control commands, and HES fire-control commands, and its reward function uses all reward terms defined in Section 3.4.

4.2 Missile air combat policy validation results

This section conducts adversarial simulations for missile-missile engagements to evaluate the missile integrated countermeasure policy. Both aircraft use identical weapon configurations, each carrying a missile. In air combat adversarial simulations, win/failure outcomes are determined by the aircraft's health state. If an aircraft is destroyed by missile attack, it is considered defeated and its opponent victorious.

The missile integrated countermeasure policy achieves convergence after approximately 1,300 training episodes. Figure 5 shows the training dynamics, including total loss and average reward.



(a) Total loss (b) Average reward

Figure 5 – Training dynamics of missile air combat policy.

The training results demonstrate a gradual decline in total loss with increasing episodes, accompanied by an overall upward trend in average reward despite periodic fluctuations and abrupt declines. This pattern indicates progressive policy refinement through iterative optimization.

A representative missile-missile engagement shows that the trained missile integrated countermeasure policy can coordinate missile launch and evasive maneuvering against the pursuit-attack policy.

4.3 HES air combat policy validation results

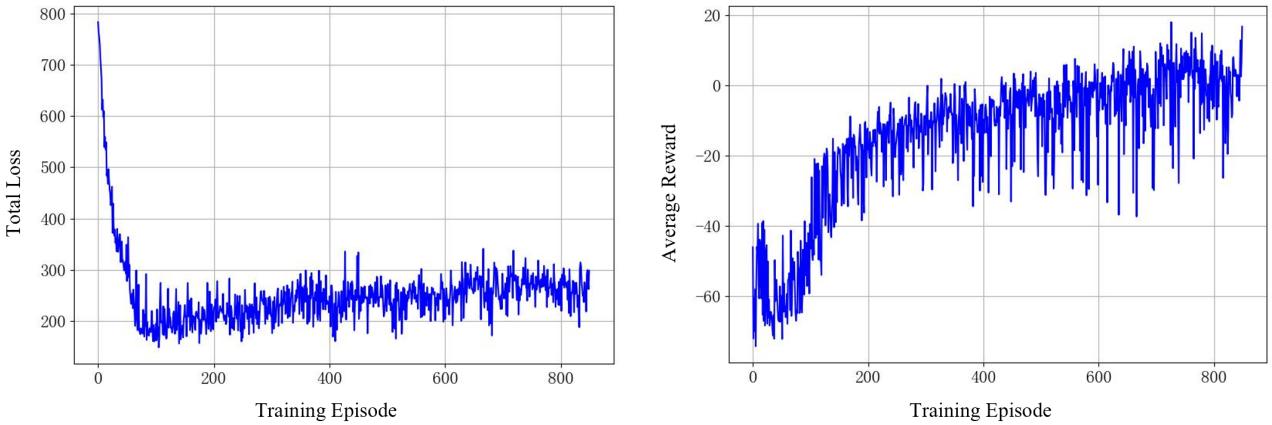
This section conducts adversarial simulations between the HES air-combat policy and the missile air-combat policy. In this configuration, the ego aircraft is equipped with both HES and missiles, whereas the enemy aircraft is equipped only with missiles. Win, failure, and draw outcomes are determined according to aircraft health states and engagement results. The ego aircraft is considered victorious if its missile damages the enemy aircraft or if the HES successfully intercepts an incoming missile.

HES integrated countermeasure policy converges after about 900 training episodes. Figure 6 shows the training dynamics, including total loss and average reward.

The training results reveal a progressive decline in total loss alongside increasing episodes, accompanied by an upward trend in average reward despite intermittent oscillations and sharp drops. These patterns indicate continuous refinement of the HES integrated countermeasure policy through iterative learning.

A representative HES engagement shows that the trained HES policy can combine missile attack with laser-based anti-missile countermeasure under the fire-control and thermal constraints.

Figure 7 shows the variation of η_{cool} during a representative HES anti-missile engagement. After HES activation at 24 s, the heat-sink residual capacity decreases and then recovers during the cooling phase, confirming the need to include thermal state in observation and fire-control logic.



(a) Total loss (b) Average reward

Figure 6 – Training dynamics of HES air combat policy.

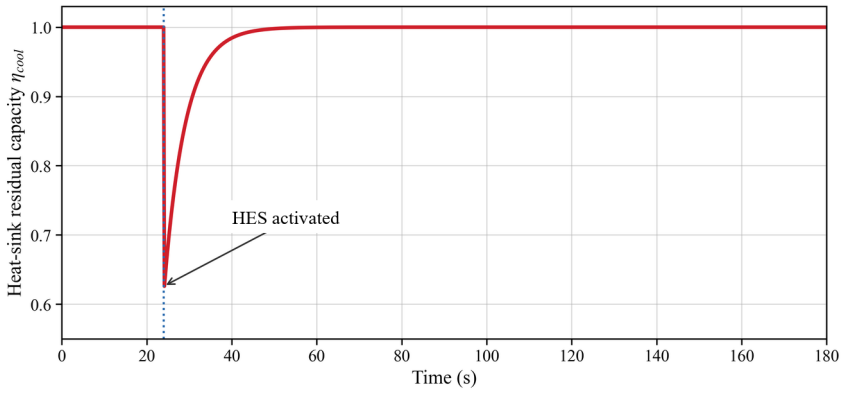


Figure 7 – Representative variation of heat-sink residual capacity during HES activation.

To evaluate the operational effectiveness of the trained HES integrated countermeasure policy, we conducted batch simulations with 100 randomized initial states within the air combat simulation framework. Both red and blue aircraft are equipped with two missiles in each simulation while employing distinct countermeasure policies. The aggregate results are summarized in Table 4.

Table 4 – HES air combat policy validation results

Red policy	Blue policy	Simulation number	Win rate	Failure rate	Draw rate
Missile integrated countermeasure policy	Pursuit-attack policy	100	56%	27%	17%
HES integrated countermeasure policy	Pursuit-attack policy	100	64%	20%	16%
HES integrated countermeasure policy	Missile integrated countermeasure policy	100	55%	18%	27%

Table 4 summarizes scenario-level comparisons within the constructed simulation setting rather than statistical significance claims. Against the pursuit-attack policy, the HES integrated countermeasure policy achieves a 64% win rate, higher than the 56% obtained by the missile integrated countermeasure policy. Against the missile integrated countermeasure policy, the HES policy still achieves a 55% win rate with an 18% failure rate. These results indicate that HES adds an additional anti-missile option and can improve tactical robustness when maneuver-only evasion is constrained.

5. Conclusions

This study establishes an integrated energy-information air combat simulation environment for strategy validation and operational effectiveness evaluation of HES. The environment couples aircraft dynamics, missile engagement, radar/fire-control logic, laser damage, and HES thermal-management models, thereby allowing the control strategy to be evaluated under dynamic thermal constraints.

Based on this environment, a hierarchical reinforcement-learning control strategy is constructed for HES air combat. The high-level Actor-Critic policy generates maneuver and HES activation decisions, the low-level PID controller executes stable flight control, and the fire-control module converts weapon commands into physically feasible engagement actions. By incorporating η_{cool} into the policy state and HES reward, the proposed strategy links thermal-state evolution with fire-control decisions. The simulation results show that the HES integrated countermeasure policy achieves a higher win rate than the missile integrated countermeasure policy against the pursuit-attack policy in this simulation setting, and maintains a lower failure rate when competing with the missile integrated countermeasure policy.

The scientific significance of this work lies in providing a closed-loop method for studying how HES thermal constraints influence tactical decision-making, rather than evaluating HES effectiveness only through static damage or platform-level thermal analysis. The results suggest that HES can provide an additional option for intercepting incoming missiles when maneuver-only evasion is insufficient. However, effective HES use requires the target to satisfy the range and angular engagement gates while sufficient heat-sink residual capacity remains, so the policy must decide when to evade, when to launch missiles, and when to activate the laser.

References

[1] Zohuri B. *Directed Energy Weapons*. Springer International Publishing, Cham, 2016.

[2] Obering H T. Directed energy weapons are real . . . and disruptive. *PRISM*, Vol. 8, No. 3, pp 36–47, 2019.

[3] Gvozdich G, Weise P and von Spakovsky M. INVENT: study of the issues involved in integrating a directed energy weapons subsystem into a high performance aircraft system. *50th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, Nashville, Tennessee, paper 2012-490, 2012.

[4] Roberts R A, Donovan A, Nuzum S R and Wolff M. Impact of high energy pulsed systems on an aircraft's power and thermal management system. *57th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, San Diego, California, paper 2016-0672, 2016.

[5] Butt N, Roberts R and Patnaik S. Transient temperature effects on the optical power wavelength shift of a high-power laser system. *Applied Thermal Engineering*, Vol. 156, pp 531–540, 2019.

[6] Cheng C, Li Y and Jiao Z. Research on thermal characteristics and thermal management of airborne fiber laser. *CSAA/IET International Conference on Aircraft Utility Systems (AUS 2022)*, Nanchang, China, pp 1313–1319, 2022.

[7] Lionis A. *Experimental design of a UCAV-based high-energy laser weapon*. PhD thesis, Naval Postgraduate School, Monterey, California, 2016.

[8] De Marco A, D'Onza P M and Manfredi S. A deep reinforcement learning control approach for high-performance aircraft. *Nonlinear Dynamics*, Vol. 111, No. 18, pp 17037–17077, 2023.

[9] Pope A P, Ide J S, Mićović D, Diaz H, Twedt J C, Alcedo K, Walker T T, Rosenbluth D, Ritholtz L and Javorsek D. Hierarchical reinforcement learning for air combat at DARPA's AlphaDogfight trials. *IEEE Transactions on Artificial Intelligence*, Vol. 4, No. 6, pp 1371–1385, 2023.

[10] Selmonaj A, Szehr O, Del Rio G, Antonucci A, Schneider A and Rüegsegger M. Hierarchical multi-agent reinforcement learning for air combat maneuvering. *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp 1031–1038, 2023.

[11] Kurniawan B, Vamplew P, Papasimeon M, Dazeley R and Foale C. An empirical study of reward structures for actor-critic reinforcement learning in air combat manoeuvring simulation. In Liu J and Bailey J, editors, *AI 2019: Advances in Artificial Intelligence*, Vol. 11919, Springer International Publishing, Cham, pp 54–65, 2019.

[12] Nuzum S. Aircraft thermal management using liquefied natural gas. *Browse all Theses and Dissertations*, 2016.

[13] Planès T, Habrard V, Delbecq S, Pommier-Budinger V and Benard E. Thermal management system models for overall aircraft design. *AIAA AVIATION 2021 FORUM*, Virtual Event, paper 2021-2428, 2021.

[14] Van Heerden A S J, Judt D M, Jafari S, Lawson C P, Nikolaidis T and Bosak D. Aircraft thermal management: practices, technology, system architectures, future challenges, and opportunities. *Progress in Aerospace Sciences*, Vol. 128, article 100767, 2022.

[15] Liu X, Lin G, Guo J, Liang X, Li Z and Jin H. Design and analysis of a novel thermal management system for serving next-generation aircraft with megawatt-level heat loads. *Case Studies in Thermal Engineering*, Vol. 61, article 104884, 2024.

[16] Reed W, von Spakovsky M and Raj P. Comparison of heat exchanger and thermal energy storage designs for aircraft thermal management systems. *54th AIAA Aerospace Sciences Meeting*, San Diego, California, paper 2016-1023, 2016.

[17] Herrera L C and Tsao B H. Analysis and control of energy storage in aircraft power systems with pulsed power loads. *SAE International Journal of Aerospace*, Vol. 9, No. 1, pp 8–13, 2016.

[18] Sun H. *A Practical Guide to Handling Laser Diode Beams*. Springer Netherlands, Dordrecht, 2015.

[19] Butt N J, Roberts R A and Patnaik S S. Laser diode optical output dependence on junction temperature for high-power laser systems. *Optics & Laser Technology*, Vol. 125, article 106019, 2020.

[20] Li L, Tao J and Gao W. Experimental study on the dynamic heat transfer characteristics of a mechanically pumped two-phase cooling loop. *Frontiers in Energy Research*, Vol. 9, article 701805, 2021.

[21] Deng Z, Shen J, Gong W, Dai W and Gong M. Temperature distribution and thermal resistance analysis of high-power laser diode arrays. *International Journal of Heat and Mass Transfer*, Vol. 134, pp 41–50, 2019.

[22] Liu Y, Li Y, Cheng C and Zhao L. Optimal design of thermal management system for airborne directed energy weapons. *2024 CSAA/IET International Conference on Aircraft Utility Systems (AUS 2024)*, Xi'an, China, pp 807–814, 2024.

[23] Sharma A, Tyagi V V, Chen C R and Buddhi D. Review on thermal energy storage with phase change materials and applications. *Renewable and Sustainable Energy Reviews*, Vol. 13, No. 2, pp 318–345, 2009.

[24] Nazir H, Batool M, Bolivar Osorio F J, Isaza-Ruiz M, Xu X, Vignarooban K, Phelan P, Inamuddin and Kannan A M. Recent developments in phase change materials for energy storage applications: a review. *International Journal of Heat and Mass Transfer*, Vol. 129, pp 491–523, 2019.

[25] Guo L, Pang L, Zhao J and Yang X. Optimization of power and thermal management system of hypersonic vehicle with finite heat sink of fuel. *Energies*, Vol. 15, No. 15, article 5332, 2022.

[26] Scukins E, Klein M, Kroon L and Ögren P. BVR Gym: a reinforcement learning environment for beyond-visual-range air combat. *arXiv preprint*, arXiv:2403.17533, 2024.

[27] Sutton R S, McAllester D, Singh S and Mansour Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, Vol. 12, 1999.

[28] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint*, arXiv:1406.1078, 2014.

[29] Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O. Proximal policy optimization algorithms. *arXiv preprint*, arXiv:1707.06347, 2017.

6. Contact Author Email Address

leeyoung303@buaa.edu.cn

7. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

8. Acknowledgement

This study was financially supported by the National Natural Science Foundation of China (No. 52272409) and by the National Key Laboratory of Aircraft Integrated Flight Control (JSY 6142219202403).