communications engineering

A Nature Portfolio journal

Article

https://doi.org/10.1038/s44172-025-00556-6

Cognitive embodied learning for anomaly active target tracking

Qihui Wu, Jiahao Li, Fuhui Zhou ✉, Jiahuan Ji, Haoyang Wang, Hongtao Liang & Kai-Kuang Ma

The primary challenge in active object tracking (AOT) lies in maintaining robust and accurate tracking performance in the complex physical scenarios. Existing end-to-end frameworks based on deep learning and reinforcement learning often struggle with high computational costs, data dependency, and limited generalization, hindering their performance in practical applications. Although embodied intelligence (EI) is promising to enable agents to learn from physical interactions, it cannot tackle severe anomalies happened in the complex scenarios. In order to address this issue, here we propose a novel embodied learning method, called the Cognitive Embodied Learning (CEL), which is inspired by the dual decision-making system of the human brain. The CEL can dynamically switch between normal tracking and anomaly handling modes, supported by specialized modules including the anomaly cognition module, the rule reasoning module, and the anomaly elimination module. Moreover, we further introduce the categorical objective function to address function non-measurability and data confusion caused by severe anomalies. Extensive unmanned aerial vehicle anomaly active target tracking experiments in both simulated and real-world scenarios demonstrate the superior performance of our method. Compared to the state-of-the-art methods, the CEL achieves a 361.4% increase in the success rate and a 54.4% improvement of the task completion efficiency, which highlights the potential of CEL to advance the field of AOT and open new avenues for more robust and intelligent tracking systems in the challenging environments.

Active object tracking (AOT) is widely applied in various domains such as autonomous vehicles[1–5], security surveillance[6,7], and human-robot interaction[8,9]. Unlike traditional passive object tracking, which relies on stationary observation, AOT involves an active agent (e.g., drone, mobile robot, or autonomous vehicle) that adaptively adjusts its movement and perspective to maintain optimal observation of a target. Particularly in the field of robotic control, where precise and real-time target tracking is critical, AOT plays an indispensable role. For example, in domestic robots[9–12], AOT enables robots to track specific targets and effectively complete household tasks, significantly enhancing adaptability and operational efficiency in domestic settings. In robotic manipulation[13–15], AOT allows robots to precisely locate and track various objects, ensuring dynamic adjustments of posture and movement trajectories during tasks like grasping and transporting, avoiding misoperation and significantly improving productivity. The existing methods for achieving AOT mainly focus on end-to-end frameworks driven by reinforcement learning and deep learning[16]. These methods integrate techniques such as adversarial learning[4,17,18], multimodal perception[19,20], self-supervised feature learning[21], active sample selection[22] and adaptive mechanism, training models to generate control actions directly from visual inputs and optimize target tracking policy. However, several factors limit the wide applicability of these end-to-end learning-based methods, such as high computational costs[23], heavy reliance on training data[24], and insufficient generalization capabilities[25]. The lack of generalization is a critical bottleneck, since policies trained for specific scenarios typically fail catastrophically when deployed in new scenarios with different visual characteristics and target behaviors. Moreover, these methods are constrained in practically complex scenarios such as cluttered urban areas with dense obstacles and rapid target maneuvers[26,27], since the learned policies often fail to adapt to unseen scenarios and require extensive retraining for each new environment or target type.

Embodied intelligence (EI) provides a new insight and solution for the implementation of AOT, especially for complex scenarios. As a representative of behaviorism, EI emphasizes the importance of the perception-action loop in the cognition[28,29]. For AOT tasks, EI enables the agent to optimize its target tracking policy via embodied actions and learn from physical interaction experiences[30–32]. For example, with the integration of EI, autonomous aerial robots and surface vehicles can adapt their motion strategies to reduce tracking errors, achieve more accurate target localization, and maintain persistent tracking of different task

College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, P. R. China. ✉e-mail: zhoufuhui@ieee.org

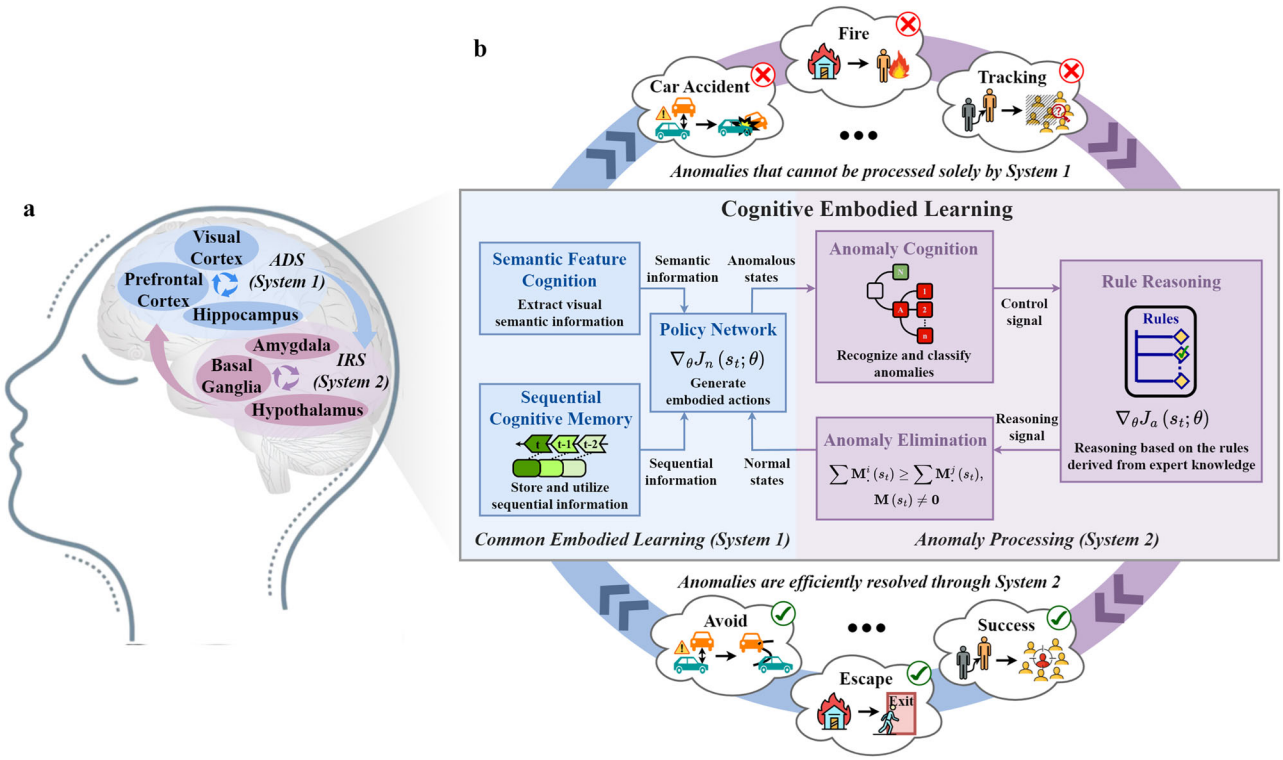

Fig. 1 | The human brain mechanism and its inspiration for our proposed CEL. a, b display the general CEL framework inspired by the dual decision-making system in the human brain. ADS is the analytic decision system (System 1), and IRS is the intuitive reasoning system (System 2).

targets[33–35]. Moreover, by incorporating physical constraints into AOT, EI can enhance the agent's understanding of the physical environment, mitigate unreasonable motion predictions, and improve the tracking ability of dynamic objects[36]. However, despite these advancements of EI in AOT, a major challenge that remains unsolved is to tackle the severe anomalies in extreme conditions. Owing to the dynamic property of the task target in AOT tasks, severe anomalies typically manifest, such as the prolonged occlusion[6,7,37–39] and intense interference[4,5,17,18], which cause the training divergence of learning algorithms and the test failure. Achieving efficient target tracking under these conditions remains a critical issue that needs to be addressed.

To tackle the aforementioned issues, we propose a novel embodied learning method called cognitive embodied learning (CEL). This method is inspired by the dual decision-making system of the human brain proposed by cognitive psychologist Daniel Kahneman[40–42], which can use two different systems to make optimal decisions based on the physical situations[43,44]. Specifically, the CEL exhibits adaptive switching between the common embodied learning mode and the anomaly handling mode in response to the normal and anomalous states. The anomaly handling mode consists of three modules—the anomaly cognition module (ACM), the rule reasoning module (RRM), and the anomaly elimination module (AEM). Moreover, inspired by the human brain's inherent ability for categorical perception during the decision-making process[45,46], we introduce the categorical objective function (COF). The COF systematically constructs different objective functions for both normal states and the different anomalies. This method addresses the issues of function non-measurability and data confusion caused by the severe anomalies[47,48], enabling the result to converge to a feasible solution.

To rigorously validate our proposed method, we choose to investigate a particularly challenging and classical task, namely, the unmanned aerial vehicle (UAV) active target tracking[49–53], which is of crucial importance in many applications. Through meticulously crafted experiments conducted in both simulated and real-world scenarios, we validate the effectiveness of our method, confirming its ability and potential to resolve severe anomalies. Compared to the state-of-the-art methods, our method achieves a 361.4% higher success rate, while improving the task completion efficiency by 54.4% under the premise of task accomplishment.

Results

Framework of cognitive embodied learning

Since the current AOT methods cannot well address the severe anomalies, they are subject to the training divergence and the test failure. Interestingly, humans can well address the severe anomalous situations due to the existence of two distinct decision-making systems, namely, the analytic decision system (ADS: System 1) and the intuitive reasoning system (IRS: System 2)[40–42]. Figure 1a illustrates the schematic overview of the dual decision-making system of the human brain. In normal situations, the human brain tends to evoke System 1 for more rational, logical, but slower decision-making. The visual cortex initially extracts semantic information from the visual information, and the prefrontal cortex, which plays a pivotal role in cognitive control, subsequently utilizes this semantic information along with the cognitive memory stored in the hippocampus to make a decision. Once a severe anomalous situation occurs, the human brain swiftly transitions from System 1 to System 2 for rapidly making a decision. The amygdala recognizes the current severe anomaly, while the hippocampus retrieves the memory to determine the danger level of the situation and utilize experience, knowledge and internalized rules to guide the response[54,55]. The basal ganglia generate habitual behaviors through the conditioned reflexes for the severe anomalous state. Once the threat is eliminated, the hypothalamus discontinues the stress response, allowing the brain to revert to System 1. These systems interact and switch dynamically during the decision-making process based on the experienced situations[40,43,44]. Inspired by these traits, we propose CEL, a novel EI method that imitates this remarkable capability of the human brain for tackling the anomalies in the AOT.

As shown in Fig. 1b, the common embodied learning mode comprises the semantic feature cognition (SFC), the sequential cognitive memory (SCM), and the policy network. The anomaly handling mode consists of

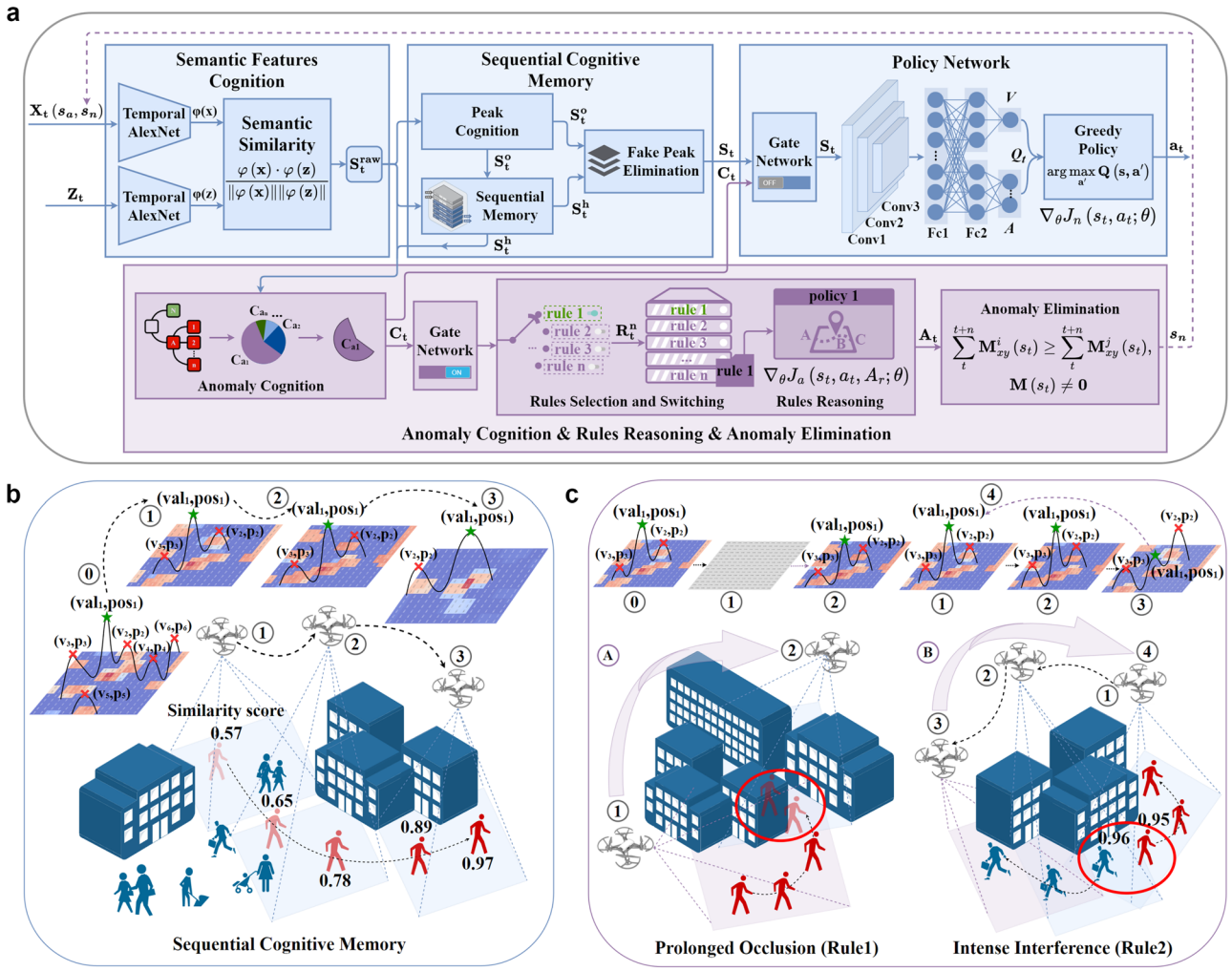

Fig. 2 | The framework and mechanism of CEL. a Implementation details of CEL in the anomaly active target tracking task. **b** The detailed processes of the sequential cognitive memory (SCM), the anomaly cognition module (ACM) and the rule reasoning module (RRM) assist the agent to recognize, categorize and resolve anomalous states. **c** Schematic for the severe anomalies of prolonged occlusion and intense interference, along with the corresponding rules-based reasoning process.

three specialized modules, namely, the anomaly cognition module (ACM), the rule reasoning module (RRM), and the anomaly elimination module (AEM), which mimics the human brain process of analyzing and handling anomalies. Specifically, in the normal state, the SFC transforms raw visual observation data into a semantic space for extracting the robust features, which is similar to the human brain's visual cortex. Then, the policy network exploits the processed semantic information along with the sequential information from the SCM to select actions based on the current state. Moreover, when severe anomalies occur, such as a car accident, a fire disaster, or prolonged occlusion and intense interference during target tracking, the ACM is responsible for recognizing and classifying severe anomalous states. Meanwhile, the SCM maintains a sequential memory to integrate multi-view, multi-scale historical sequential information to guide the cognition of severe anomalies. Then, the RRM retrieves the corresponding rules, which are derived from the structured expert-embedded knowledge, to devise a recovery policy for reasoning. Finally, the AEM evaluates the current state and eliminates the severe anomalies, thus restoring the framework to the normal state and the corresponding decision-making mode. To evaluate the effectiveness of each module, we conduct detailed ablation experiments (see Supplementary Note 9 for details).

The operational flow of the CEL framework can be summarized and shown in Fig. 2a. Firstly, the SFC processes the raw data X_t for extracting the semantic information, which is represented as the current state S_t^{raw}. Next, S_t^{raw} is inputted into both the SCM and the ACM for further processing. In the SCM, it processes S_t^{raw} into the new current state S_t, which is a semantic similarity map. Then, S_t^{raw} and the historical sequential information S_t^h are jointly input into the ACM, serving as the overall state information to assist the ACM in recognizing whether the CEL agent experiences severe anomalies or not, categorizing the types of severe anomalies, and ultimately generating the control signal C_t. If the CEL agent does not trap into a severe anomaly, the control signal C_t is set as 0, and guides S_t to flow into the policy network to generate the action a_t. Otherwise, the RRM exploits rules based on the cognition results of the ACM to reason the action sequence A_t for the corresponding severe anomalous states. Finally, the AEM evaluates the current state and eliminates the severe anomalies based on the reasoning results, thus restoring the framework to its decision-making mode of normal states.

The CEL framework for the anomaly active target tracking task

Anomaly active target tracking is a practical task subject to severe anomalies. The severe anomalies happen during the tracking process due to the inherent dynamic property of the task target. Those severe anomalies include prolonged occlusion caused by building obstructions or the task target's intentional evasion, as well as intense interference resulting from the task target deliberately adopting highly similar appearance characteristics to those of distractors in the environment. Therefore, we consider this challenging and illustrative task to validate the efficacy of the proposed CEL for

handling severe anomalies. In this task, we employ a UAV to sustainably track the task target, solely relying on the raw observation images and the provided target template images. To handle severe anomalies, such as prolonged occlusion and intense interference, rules derived from the structured expert-embedded knowledge are used for reasoning. Once severe anomalous states are recognized by the ACM, the UAV dynamically switches to the RRM, allowing it to rediscover and re-lock onto the task target after breakdown events.

The implementation details of the CEL framework for the anomaly active target tracking task is shown in Fig. 2a. The current raw observation image X_t is gridded into $n*n$ small image blocks. The SFC first extracts features from each small image block and the target template images. Then, by computing the distance between these extracted features, it obtains the semantic similarity map as the current state S_t^{raw}, i.e., the similarity between each small image block and the template (see Methods for details). The semantic feature extraction process of the SFC maps the raw data layer to a high-level semantic layer. Since the similarity value is universal across different scenarios, the policy based on it can adapt well to new scenarios.

After the semantic feature extraction process, the SCM maintains a sequential memory to integrate and utilize multi-view, multi-scale historical sequential information. S_t^{raw} is fed into the peak cognition module, which extracts the peak values and their corresponding location information in the semantic similarity map, and finally obtains the observation state S_t^o. S_t^{raw} and S_t^o are stacked together and then stored in the sequential memory module. The sequential memory module stores $k+1$ historical information as the historical sequential information S_t^h, which contains multi-view, multi-scale historical information about the task target. S_t^o and S_t^h are jointly inputted into the fake peak elimination (FPM) module. This joint input is a deliberate design where these two components have distinct and complementary roles. Specifically, S_t^o serves as the real-time query, providing the module with the latest, unprocessed peak information from the present frame, which ensures immediate responsiveness to new and rapidly changing target appearances. In contrast, S_t^h, which contains a sequence of past observations including the current state S_t^o, performs as the temporal context (i.e., keys and values in the attention mechanism). The FPM module utilizes the current state S_t^{raw}, the historical sequential information S_t^h, and target similarities at different times and locations. The similarities are fused and superimposed. By querying the current peak candidates (S_t^o) within the historical sequential context (S_t^h), the FPE module enhances the similarity value of the true task target while suppressing transient distractor interference and noise. During the process of fusion and superposition, the true peak value in the semantic similarity map gradually increases while the pseudo-peak values decrease (Fig. 2b). Eventually, a clear semantic similarity map is obtained, highlighting the true peak. This map serves as the new current state S_t. In normal states, the policy network receives S_t and selects the action a_t based on the greedy policy.

Then, the ACM assesses whether the agent experiences the severe anomalous states based on the semantic similarity matrix $\mathbf{M}(s_t)$ (see Methods for its detailed definition) or not. When $\mathbf{M}(s_t)$ is an all-zero matrix, it indicates that the agent experiences a severe anomaly due to the prolonged occlusion. Conversely, if the true peak value corresponding to the task target in $\mathbf{M}(s_t)$ suddenly drops while the pseudo-peak values corresponding to certain distractors abruptly rise, it signifies that the agent encounters a severe anomaly due to intense interference (see Methods for the detailed mathematical descriptions). If a severe anomaly is recognized, the ACM identifies the specific type of the severe anomaly, and the RRM selects the corresponding rule from the rule base for reasoning based on the cognition results of the ACM. Then, the policy A_t corresponding to the selected rule is outputted by the ACM. Finally, the AEM evaluates the current state and eliminates the severe anomalies based on the reasoning results of the RRM and the corresponding cognition conditions, thereby restoring the framework to its normal state.

In the UAV anomaly active target tracking task, the rule base of the RRM has two important predefined rules to address severe anomalous states. Specifically, Rule 1 is used to solve the severe anomaly of prolonged occlusion (Fig. 2c). When the task target is obscured by buildings or intentionally evades tracking, the UAV recognizes the current severe anomalous state and uses Rule 1 to fly directly over the obstructions and reach the opposite side to resume tracking of the task target. Rule 2 is used to solve the severe anomaly of intense interference (Fig. 2c). When the task target is interfered with by similar distractors and is eventually lost, the UAV also recognizes the anomaly and uses Rule 2 for reasoning. The UAV uses the historical sequential observation information from the SCM to immediately return to the landmark closest to the one before the task target is lost. The landmark indicates the time when the task target is last detected by the UAV. It is used to recapture the task target and resume detection and tracking. In the absence of rules, the UAV is likely to fail the tracking task, since the learning network struggles to adapt to those severe anomalies. (See Methods for details regarding Rule 1 and Rule 2).

The training process of CEL for the anomaly active target tracking task

The training process of the CEL agent for the UAV anomaly active target tracking task is shown in Fig. 3. To train the CEL agent, we build training scenarios in the Gazebo simulation environment[56], which is an open-source 3D robotics simulator featuring robust physics engines. By leveraging Gazebo's realistic sensor simulations and flexible environment customization capabilities, the intricate real-world dynamics of the target tracking task can be reproduced to efficiently train the CEL agent in a safe, accelerated and cost-effective manner. The training scenarios expose the CEL agent to prolonged occlusion caused by buildings, shadows and other obstacles, as well as intense interference from distractors that highly resemble the task target, which are very common but extremely challenging in the UAV anomaly active target tracking task.

During the training process, the CEL agent collects training data through constant interactions with training scenarios in order to train the policy network in the CEL framework. In training scenarios, the agent is concurrently presented with normal states as well as diverse severe anomalous states. At these times, the policy network and the RRM function alternately. Note that in the above process, only the data generated by the agent using the policy network for decision-making under normal states is stored in the experience replay buffer, and the trajectories generated by the RRM are not stored. The reason is that when the agent experiences prolonged occlusion or intense interference, the semantic similarity map produces anomalies. If the task target is severely occluded, it disappears from the raw observation image. The similarity value of each small image in the $n*n$ grid is lower than the threshold value, and none of them completes the similarity matching with the template images. Therefore, the semantic similarity map is an all-0 anomaly matrix. Meanwhile, if the agent experiences intense interference from a distractor and loses sight of the task target, the distractor appears on the raw observation image instead. This occurs because the matching similarity value between the distractor and the task target template images may reach its maximum value at a certain scale, viewpoint, and moment. At this point, the agent may lose track of the task target and begin to track the distractor. However, as the scale of view changes, the matching similarity value rapidly decreases. Only the real task target and the template images maintain the maximum matching similarity value over time. In the grid of each small image, the similarity value for matching with the template images is low. The highest peak value on the semantic similarity map remains in an anomalous state. Storing the anomalous states and the corresponding actions generated by the RRM in the experience replay buffer can affect the original data distribution and violate the laws in the original trajectory data. Therefore, it is important to avoid storing these states in the experience replay buffer. This data confusion can prevent the policy network from learning the optimal decision policy patterns present in the trajectory data. The specific process is shown in Fig. 3.

At each time step, the agent utilizes current and historical raw observation images to generate actions through the policy network or the RRM. The raw observation images are converted into semantic similarity maps via the SFC. The policy network and the RRM receive the semantic similarity

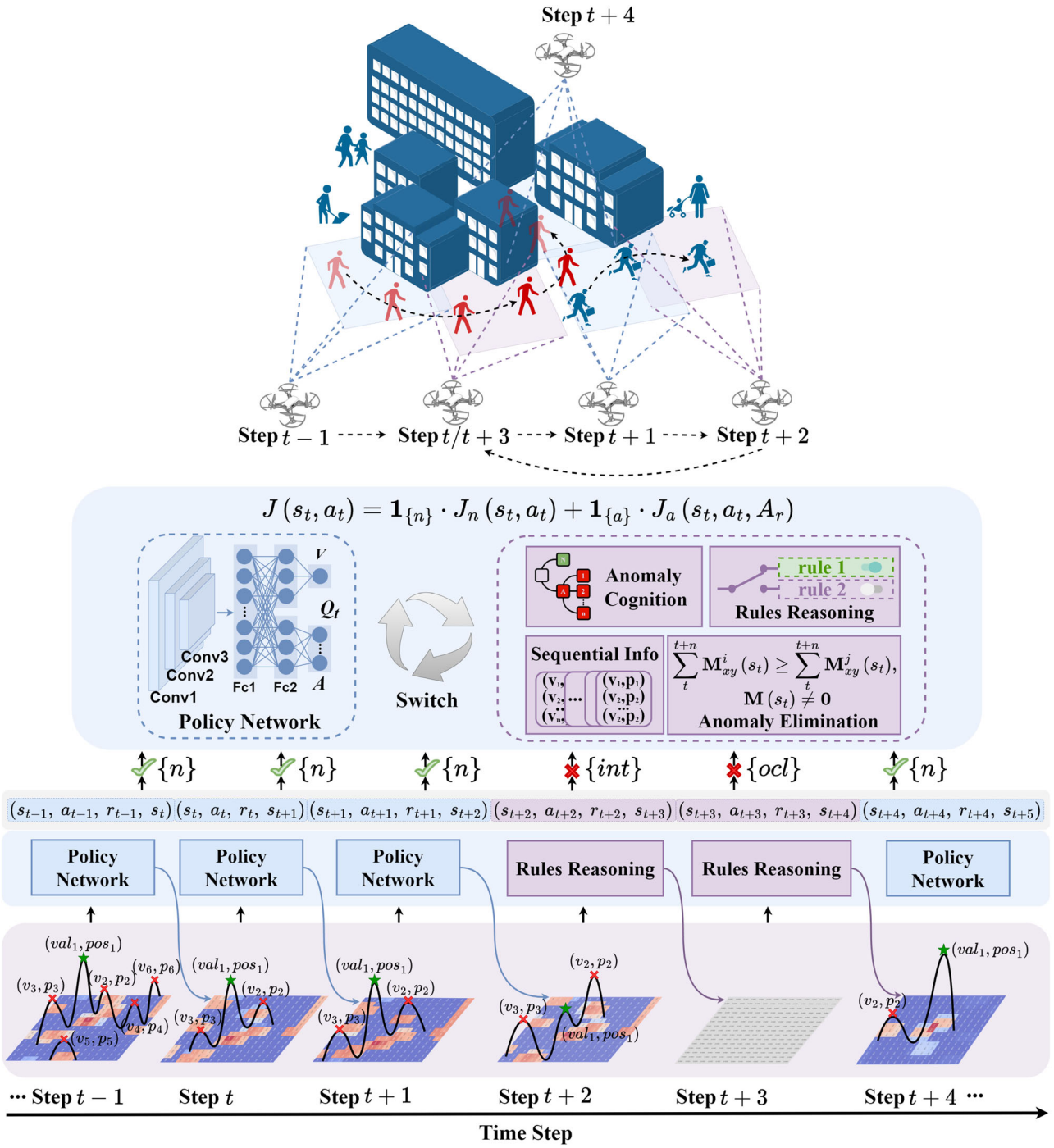

Fig. 3 | The training process of the CEL framework in the UAV anomaly active target tracking task. The vertical flow from the bottom to the top represents the whole process of the CEL agent, which includes semantic information extraction, anomaly cognition, transition between the policy network and the rule reasoning module (RRM), and the respective decision-making and reasoning. The agent can be effectively trained on a mixture of normal and severe anomalous data. It recognizes and categorizes the anomaly occurrences and types, enabling the distinction and removal of the severe anomalous data from the normal. It selects an appropriate objective function based on the cognition results to optimize the training process, preventing the training divergence due to severe anomalous data confusion.

maps, along with the information related to the current position, the historical positions, and the semantic similarity maps themselves, which together constitute the current state for decision-making. The RRM operates without the need for training. It leverages the cognition results of the current state through the ACM and employs predefined rules derived from the structured expert-embedded knowledge for reasoning. In contrast, the policy network requires training and is updated through the collected experiences (s_t, a_t, r_t, s_{t+1}), which is gathered during interactions of the agent with the training scenarios.

The current state s_t is defined as

$$s_t = [s_t^o, s_t^h, p_t, p^h, m_t, m^h], \tag{1}$$

where s_t^o and s_t^h denote the current and historical observation state, respectively, while p_t and p^h represent the current and historical position states, and m_t and m^h signify the current and historical information about the semantic similarity maps. $p_t = [x_t, y_t, z_t, \theta_t]$, where x_t, y_t, and z_t coordinates stand for the current position of the agent, while the rotation angle of

its view field is represented by θ_t. $m_t = [n, v_t^1, x_t^1, y_t^1, \ldots, v_t^n, x_t^n, y_t^n]$, where n indicates the number of peaks in the semantic similarity map. v_t^i denotes the peak value of the ranked i-th at time t, while x_t^i and y_t^i represent the coordinates of its location. All historical state information, such as s_t^h, p^h, and m^h, contains the states of the last five consecutive time steps.

The action a_t is defined as the flight direction. The discrete action space A is represented by a 12-dimensional vector containing flying up and down, left and right, front and back, upper left and upper right, lower left and lower right, rotate left and rotate right. The reward r_t is defined as

$$r_t = r_t^g + r_t^m + r_t^l + r_t^b, \tag{2}$$

where r_t^g is the reward for tracking the task target, r_t^m is the reward about the semantic similarity map, r_t^l is the penalty for losing the task target and r_t^b is the penalty for reaching the boundary. In normal states, r_t serves as the reward function for training the policy network. However, in severe anomalous states, the RMM directly performs reasoning without requiring training, and thus, r_t only reflects the rules-based reasoning performance of the RMM.

r_t^m is a metric that reflects the quality of the current observation semantic similarity map. The semantic similarity map reward r_t^m is defined as

$$r_t^m = \frac{1-n}{N}\sum_{i=2}^{n}(v_t^1 - v_t^i), \tag{3}$$

where n, v_t^1 and v_t^i are consistent with their definitions in m_t, N denotes the total number of peaks in the semantic similarity map. m_t aims to encourage the agent to generate trajectories with distinct semantic similarity, which is achieved by increasing the difference between the true and pseudo peaks while reducing the number of pseudo peaks.

During the tracking process, the CEL agent recognizes the normal and severe anomalous categories based on their characteristics. In the event that the CEL agent identifies itself within a severe anomalous state, it proceeds to further classify this state into either a prolonged occlusion or an intense interference.

In normal states, the CEL agent focuses on optimizing the objective function $J_n(s_t, a_t)$. Conversely, when encountering severe anomalous states, the CEL agent switches its attention to optimize the objective function $J_a(s_t, a_t, A_r)$. Moreover, the agent faces the prolonged occlusion anomaly state, it shifts its optimization towards the objective function $J_{ocl}(s_t, a_t, A_r)$, while in the intense interference anomaly state, it directly optimizes the objective function $J_{int}(s_t, a_t, A_r)$. The ensemble of these objective functions constitutes our proposed categorical objective function (COF), which serves as the mathematical foundation for the CEL framework (see Methods for details). The COF achieves fitting of unmeasurable functions during the training process by switching among different objective functions under normal and varying severe anomalous states, thereby addressing the training divergence issue under severe anomalies.

Experiment setting

We conduct extensive UAV anomaly active target tracking experiments in both the simulated and physical scenarios to validate the superior performance of the proposed CEL framework in resolving the training divergence and the test failure caused by the severe anomalies. Five state-of-the-art benchmark methods are compared with the CEL.

Regarding the CEL method, CEL(Rule1), CEL(Rule2), and CEL(Rules) respectively indicate that the RRM reasons with the rule base containing Rule 1, Rule 2, or their combination. Meanwhile, EL(EA + SI), EL(EA), EL(SI), and EL(STF) represent four common embodied learning paradigms. The EA + SI method employs the SFC, SCM, and the policy network within the CEL. Raw observation images are processed by the SFC and SCM to inform the policy network's final decision, without involving the RRM. The EA method solely uses the SFC and the policy network in the CEL. The policy network receives the raw observation images as decision inputs after processing by the SFC. The SI method uses the SFC and SCM to process each frame of the raw observation image into a semantic similarity map, termed the detection token. This token is then fed into a transformer instead of the policy network. After training, the transformer generates policy tokens that predict the agent's next action, subsequently deriving the action probability distribution via the MLP. Moreover, the STF method employs a pioneering adaptive embodied learning framework designed for interactive environments[18,57]. It draws inspiration from the human tutorial learning process and involves a student network learning under the guidance of a teacher network that evaluates predictions. The framework demonstrates strong generalization across diverse environments and effectively resolves occlusion and interference. Notably, distinct from the aforementioned methods, the NBV[58] is a computational approach that employs a region-based analysis of an initial target observation image to determine optimal subsequent views based on the model. In the task of anomaly active target tracking, it can suggest the next viewpoint based solely on the information derived from a single initial observation view without the need of a predefined dataset.

The training results of diverse methods across various scene settings are initially presented, followed by a thorough assessment of the tracking performance. Subsequently, the CEL proficiency in rules-based decision-making and its overall generalization performance are rigorously evaluated.

Overcoming the training divergence with the CEL

In four different training scenarios, we train seven different agents: CEL(-Rules), CEL(Rule1), CEL(Rule2), EL(EA + SI), EL(EA), EL(SI), and EL(STF). The results are shown in Fig. 4. Specifically, in an episode with a limited number of steps, the agent begins the task from a randomly assigned initial position. Upon detecting the task target, the agent receives a corresponding training reward and continues to track the task target until it either reaches the maximum number of episode steps or loses the task target. Subsequently, the agent reinitiates the task with a new randomized starting point and a new task target. The training evaluation metric is the average episode reward, which is calculated as the average of the episode cumulative rewards over the last 1000 episodes. A higher reward signifies more instances of the successful task target detection, prolonged tracking of the task target, and increased effectiveness of the learned tracking policy.

Figure 4a illustrates the training performance of different agents in Training Scenario-1. This scenario represents a training environment with a mixture of severe anomalies, including both prolonged occlusion and intense interference, which imposes an extreme challenge to the agents. As depicted in Fig. 4a, the training process of CEL(Rules) does not display any divergence, and the agent ultimately attains an average episode reward of around 135. Benefiting from the advantages of the CEL framework, the CEL(Rules) agent exhibits cognition of diverse anomalies encountered during training and utilizes the policy network or RRM to make decisions under both normal and different anomalous states. Moreover, it dynamically selects different objective functions from the COF based on the current state, thereby enhancing the stability of the overall training process and overcoming the issue of training divergence caused by the severe anomalies. Meanwhile, the combination of Rules 1 with 2 enables the CEL(Rules) agent to effectively address prolonged occlusion and intense interference. As a result, the CEL(Rules) agent outperforms other agents, achieving the best training performance. For the CEL(Rule1) and CEL(Rule2) agents, the training divergence is not happened. However, due to their utilization of distinct rules, they exhibit different training performance with the average episode rewards of ~122 and 106, respectively. In the anomaly active target tracking task, prolonged occlusion tends to exert a more significant adverse impact on the agent during the training process compared to intense interference. Consequently, the CEL(Rule1) agent, which employs Rule 1 to address the severe anomaly of prolonged occlusion, outperforms the CEL(Rule2) agent. However, its efficacy remains inferior to that of the CEL(Rules) agent, which leverages the combination of rules. For the EL(EA + SI), EL(EA), and EL(SI) agents, which do not leverage any rules, struggle to cope with the severe anomalies. Their training processes exhibit

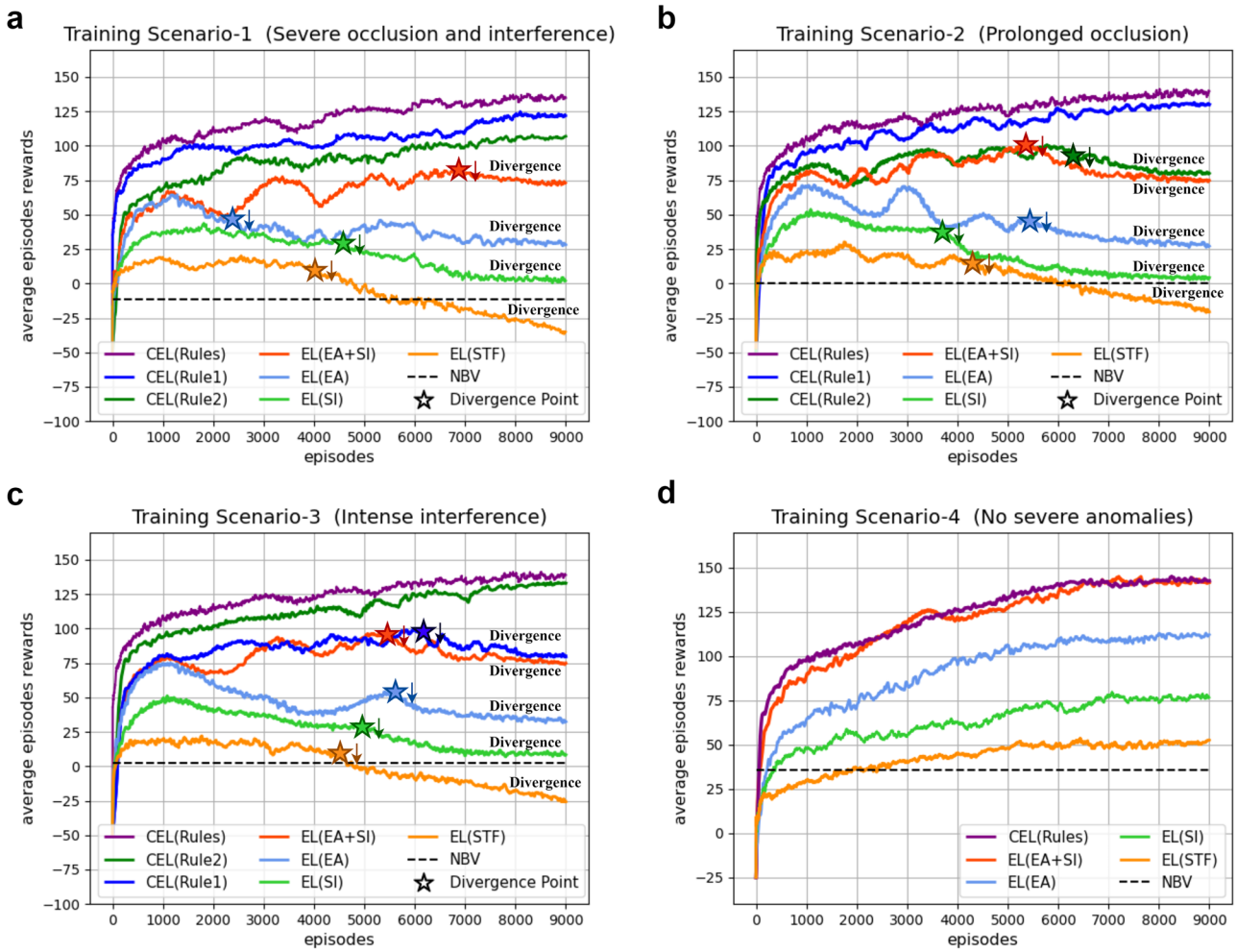

Fig. 4 | Training results of different methods in various training scenarios. **a**–**c** Training results of different methods in training scenarios with severe anomalies involving prolonged occlusion, intense interference, and their combination. Effective cognition of anomalies and the use of appropriate rules for anomaly handling can significantly enhance the algorithm training effect and address the issue of training divergence. **d** Training results of different methods in the anomaly-free training scenario. In the absence of anomalies, rules do not exert their corresponding functions, but our method continues to exhibit superior performance. EL(EA + SI), EL(EA), EL(SI), and EL(STF) represent four common embodied learning (EL) paradigms. EA+SI is a policy learning method that uses the semantic feature cognition (SFC) and the sequential cognitive memory (SCM) to process observations. EA is a policy learning method that uses the SFC to process observations. SI is a transformer-based method that uses the SFC and the SCM to generate semantic tokens. STF is an adaptive embodied learning method based on the student-teacher framework. NBV is a region-based viewpoint optimization computational method.

varying degrees of divergence at approximately 6800, 2400, and 4600 episodes. Notably, without the strengths of sequential information and embodied action, the training divergence trend is more serious in the EL(EA) and EL(SI) agents. The performance of EI(STF) and NBV agents is even worse. The NBV agent, which relies on the computational approach, attains an average episode reward of −11, while the EI(STF) agent encounters the divergence around 4000 episodes, with the average episode rewards ultimately diverging to approximately −40.

To further investigate the specific performance of Rule 1 and Rule 2 in the scenarios with the prolonged occlusion and intense interference, we train all the aforementioned agents in Training Scenario-2 and 3. In Training Scenario-2, the severe anomaly of prolonged occlusion is primarily present. As illustrated in Fig. 4b, all agents except for CEL(Rules) and CEL(Rule1) experience the training divergence. The CEL(Rule1) agent achieves comparable training performance to that of the CEL(Rules) agent. They attain average episode rewards of around 137 and 131. In contrast, the CEL(Rule2) agent demonstrates training performance similar to that of the EL(EA + SI) agent without any rules. They display the training divergence at ~6300 and 5400 episodes. Similarly, as observed in Fig. 4c, in the Training Scenario-3, which is mainly characterized by intense interference, only the CEL(Rules) and CEL(Rule2) agents avoid the training divergence. The performance of the CEL(Rule2) and CEL(Rule1) agents closely approximates that of the CEL(Rules) and EI(EA + SI) agents. These results indicate that only Rule 1 demonstrates effectiveness in Training Scenario-2, whereas the situation is reversed in Training Scenario-3. This demonstrates that different rules are effective only for their corresponding anomalous states. To address various abnormal conditions, it is necessary to formulate appropriate rules based on the specific characteristics of the anomalies and domain expert knowledge.

As depicted in Fig. 4d, in Training Scenario-4, where no anomalies are present, none of the agents exhibit the training divergence. This further indicates that the severe anomalies are the underlying reason of the training divergence. Meanwhile, the training performance of CEL(Rules) and EL(EA + SI) agents is essentially consistent. This observation suggests that although the rules do not exert positive influences in the anomaly-free scenario, concurrently, owing to the anomaly cognition mechanism of the CEL framework and the corresponding modules transition mechanism between the normal and diverse anomalous states, the presence of rules does not adversely affect the training process under normal states.

The test failure mitigation and the related performance analysis

To evaluate the anomaly active target tracking performance of all agents, we conduct tests in Test Scenario-1. This scenario stochastically introduces occlusions of the task target and generates a variable number of distractors to interfere with the target tracking. The placement of occlusions and distractors is also randomized. Moreover, to quantitatively assess the tracking capabilities of all trained agents, we employ two evaluation metrics, namely, the success rate (SR) and the relative path length (RPL), to measure their performance. The SR reflects the degree of the test failure encountered by the agent during the actual execution of tracking tasks. A higher SR indicates fewer test failures, while a lower SR corresponds to more frequent failures. The RPL reflects the agent tracking efficiency in successfully completing the task. A higher RPL indicates lower tracking efficiency, and vice versa. (See Methods for their detailed definitions).

As shown in Fig. 5a, b, the x-axis denotes the maximum allowed steps for the evaluation episode, during which a trained agent tracks from a randomized initial position to the task target. The episode terminates either when the agent loses sight of the task target or when the predefined step limit is reached. In Testing Scenario-1, the CEL(Rules) agent achieves superior performance, attributed to its proficient anomaly cognition and judicious rules utilization. This capability enables it to effectively address prolonged occlusion, intense interference and the resulting test failure. It achieves the highest SR and the shortest RPL across varying step limits compared to other agents. When the maximum episode length is constrained to 3000 steps, the CEL(Rules) agent achieves an SR of 95.5% and an RPL of 6.7. The SR surpasses that of the EL(EA + SI) agent by 16.9%, and demonstrates an even more significant improvement of 361.4 and 640.3% compared to the EL(STF) and EL(NBV) agents. This suggests that the CEL(Rules) agent encounters fewer test failures during the actual task execution. Meanwhile, in terms of efficiency, the CEL(Rules) agent obtains a 21.2% reduction of RPL relative to the EL(EA+SI) agent, and further reductions of 54.4 and 56.8% compared to the EL(STF) and EL(NBV) agents. These empirical results indicate that the CEL(Rules) agent substantially outperforms other benchmark agents in terms of the task completion efficiency, while effectively mitigating the test failure.

Notably, as the maximum allowed steps increase, both SR and RPL of all agents exhibit upward trends. However, the CEL(Rules) agent shows a relatively modest rise of RPL despite the increase in SR. Moreover, when the maximum number of steps is constrained, such as a setting of 1000 steps, which simulates realistic conditions with limited time for target detection and tracking, the CEL(Rules) agent demonstrates significant advantages in both SR and RPL over other agents. These findings highlight the CEL(Rules) agent's proficiency in learning a highly accurate and effective target tracking policy.

Rules performance evaluation

To comprehensively evaluate the performance of different rules under various severe anomalies, we further conduct tests in Test Scenario-2 and 3. In Test Scenario-2, we use the prolonged occlusion as the severe anomaly. As illustrated in Fig. 5c, d, the x-axis represents varying occlusion ratios (from 40 to 70%) for the task target. These ratios correspond to different durations of the target occlusion. A higher occlusion ratio indicates more significant occlusion and longer duration during which the target is occluded, thereby increasing the difficulty for agents to complete the tracking task. Our analysis of Test Scenario-2 reveals a notable performance pattern when prolonged occlusion is the primary severe anomaly. Across varying occlusion ratios, the CEL(Rule1) and CEL(Rules) agents, employing Rule 1 alone or in conjunction with Rule 2, consistently achieve optimal performance. Specifically, when the occlusion ratio reaches 70%, they attain a superior SR of 80.2 and 80.8%, and a minimal RPL of 12.1 and 11.8, respectively. Meanwhile, since Rule 2 is tailored for intense interference and ineffective against prolonged occlusion, the CEL(Rule2) agent's performance is comparable to that of the EL(EA + SI) agent without any rules. The SR of CEL(Rule2) agent is 63.7%, which is 20.6% lower than that of the CEL(Rule1) agent. The RPL is 18.7, 54.5% higher than that of the CEL(Rule1) agents.

In Test Scenario-3, intense interference is the other severe anomaly. As depicted in Fig. 5e, f, the x-axis represents the number of distractors (from 1 to 10) that interfere with the task target during the task. The number of distractors serves as the indicator of the degree of interference with the task target. Similarly, in Test Scenario-3, the CEL(Rule2) and CEL(Rules) agents, utilizing Rule 2 alone or in combination with Rule 1, consistently achieve prime SR and RPL across varying numbers of distractors. In the most severe interference setting (with ten distractors), they attain SR of 82.8 and 83.8%, and RPL of 9.1 and 8.8, respectively. In contrast, as Rule 1 is only applicable to prolonged occlusion and not to intense interference, the performance of the CEL(Rule1) agent is similar to that of the EL(EA+SI) agent. The SR of CEL(Rule1) agent is 70.8%, which is 14.5% lower than that of the CEL(Rule2) agent, and its RPL is 11.0, which is 20.9% higher than that of the CEL(Rule2) agent.

Notably, in both Test Scenario-2 and 3, the performance of the hybrid-rules agent (CEL (Rules)) is slightly superior to that of single-rule agents (CEL(Rule1/2)). This is attributed to the fact that neither simulation nor real-world test environments can effectively isolate a single severe anomaly alone. Instead, one severe anomaly typically predominates while others may also occur with a certain probability (e.g., occlusion often accompanies crowd interference). In those realistic conditions, the CEL framework can well recognize and handle severe anomalies.

Generalization performance evaluation

To further assess the generalization performance of the CEL framework, we introduce another two new test scenarios, Test Scenario-4 and Test Scenario-5. In Test Scenario-4 and 5, the agents are tasked with target tracking in a broader range compared to Test Scenario-2 and 3. Additionally, the arrangements of Test Scenario-4 and 5 are novel and distinct from all previous training and testing scenarios. We evaluate the performance of all agents in these new test scenarios, and the results are presented in Tables 1 and 2.

Table 1 illustrates the anomaly active target tracking performance of all agents in Test Scenario-4. As the number of distractors increases, the performance of the EL(STF) agent is obviously decreased, approaching that of the NBV agent. This decrease is due to the heavy reliance of the learned policy on the features present in the scenario. Since the feature distribution in Test Scenario-4 differs significantly from that of the training scenarios, the learned policy proves ineffective in this new scenario. Simultaneously, the performance of the EL(SI), EL(EA), and EL(EA + SI) agents that do not employ rules also deteriorates remarkably, with a notable decrease in SR and a substantial increase in RPL. Notably, when the number of distractors reaches 10, the NBV agent exhibits severe test failures and is entirely incapable of completing the target tracking task.

Meanwhile, Table 1 also indicates that the CEL(Rules) agent outperforms all other agents in SR and RPL. In Test Scenario-4, when confronted with ten distractors, the SR of the CEL(Rules) agent reaches 76.9%, whereas the EL(STF) agent achieves only 17.1%. These findings highlight the CEL(Rules) agent's proficiency in efficiently utilizing rules to address severe anomalies in new scenarios. Moreover, the CEL(Rules) agent makes decisions based on multi-view, multi-scale historical sequential semantic information, enabling its learned policy to track targets effectively even in an unfamiliar context. This emphasizes the efficacy of the CEL framework in enhancing agents' generalization abilities for novel environments.

Table 2 presents the anomaly active target tracking performance of all agents in the more challenging Test Scenario-5. This scenario poses extreme challenges due to the scenario being entirely new, random prolonged occlusion occurring, and agents cannot predict that the occlusion happens, thereby complicating the task of target tracking.

Comparing the results in Table 1, the SR and RPL of all agents show varying degrees of decrease and increase. Remarkably, the CEL(Rules) agent consistently outperforms all other agents. This suggests that the CEL framework is generalized to complex scenarios. Moreover, the performance gap between the CEL(Rules) agent and other agents increases. For instance, as shown in Table 2, in Test Scenario-5 where the occlusion ratio is 70%, the

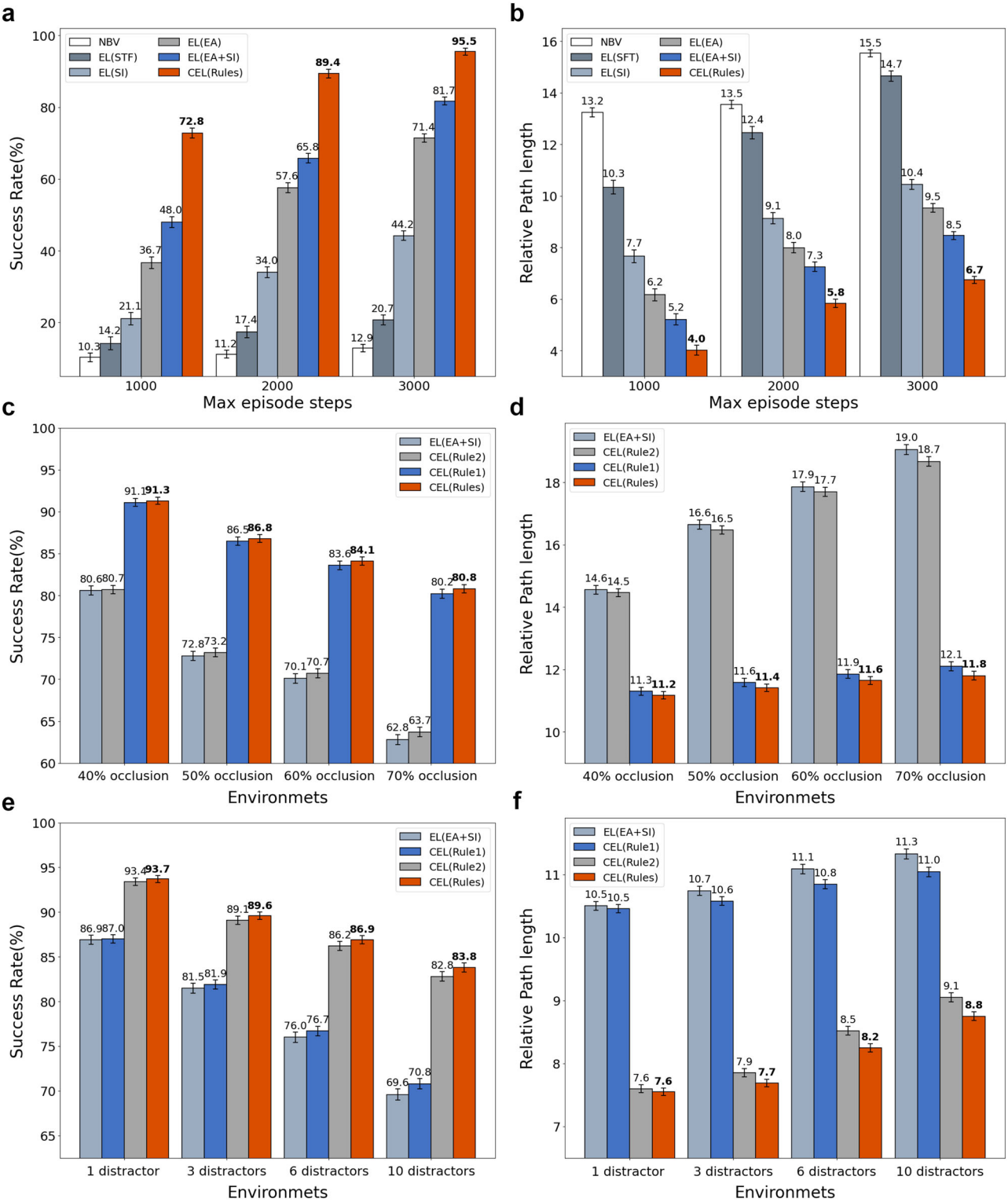

Fig. 5 | Anomaly active target tracking performance of all agents in different test scenarios. a, **b** display the success rate (SR) and the relative path length (RPL) for different agents in the Test Scenario-1. **c**–**f** display variations of SR and RPL across test scenarios-2 and test scenarios-3 among agents using rules versus those without, and among agents employing different rules. EL(EA + SI), EL(EA), EL(SI), and EL(STF) represent four common embodied learning (EL) paradigms. EA + SI is a policy learning method that uses the semantic feature cognition (SFC) and the sequential cognitive memory (SCM) to process observations. EA is a policy learning method that uses the SFC to process observations. SI is a transformer-based method that uses the SFC and the SCM to generate semantic tokens. STF is an adaptive embodied learning method based on the student-teacher framework. NBV is a region-based viewpoint optimization computational method. Error bars represent standard deviation (SD) across ten independent runs. Each run corresponds to an evaluation with a different random seed.

Table 1 | Generalization performance evaluation in Test Scenario-4

Method	No distractor		1 distractor		3 distractors		6 distractors		10 distractors	
	SR(%)	RPL	SR(%)	RPL	SR(%)	RPL	SR(%)	RPL	SR(%)	RPL
NBV	35.6	25.5	18.8	51.8	10.9	102.4	5.3	119.2	-	-
EI (STF)	54.8	15.8	43.7	30.1	35.6	49.8	26.4	101.4	17.1	125.8
EI (SI)	69.5	15.7	62.0	19.7	56.0	22.6	50.0	26.3	41.7	30.1
EI (EA)	77.9	14.3	70.6	15.4	64.8	16.0	57.4	16.7	49.6	17.4
EI (EA+SI)	81.4	12.1	75.8	12.8	71.7	13.2	66.5	13.7	60.4	14.3
CEL (Rules)	**91.1**	**8.1**	**89.5**	**8.6**	**85.6**	**8.9**	**81.2**	**9.5**	**76.9**	**10.1**

The values in bold indicate the best-performing results in every setting. SR is the success rate, and RPL is the relative path length. EL(EA + SI), EL(EA), EL(SI), and EL(STF) represent four common embodied learning (EL) paradigms. EA+SI is a policy learning method that uses the semantic feature cognition (SFC) and the sequential cognitive memory (SCM) to process observations. EA is a policy learning method that uses the SFC to process observations. SI is a transformer-based method that uses the SFC and the SCM to generate semantic tokens. STF is an adaptive embodied learning method based on the student-teacher framework. NBV is a region-based viewpoint optimization computational method.

Table 2 | Generalization performance evaluation in Test Scenario-5

Method	30% occlusion		40% occlusion		50% occlusion		60% occlusion		70% occlusion	
	SR(%)	RPL	SR(%)	RPL	SR(%)	RPL	SR(%)	RPL	SR(%)	RPL
NBV	2.8	105.7	1.9	117.5	-	-	-	-	-	-
EI (STF)	9.7	56.3	2.7	97.4	-	-	-	-	-	-
EI (SI)	18.9	21.7	10.6	49.6	3.8	108.3	-	-	-	-
EI (EA)	66.8	15.1	60.9	18.9	52.7	21.8	45.4	24.6	38.0	28.3
EI (EA+SI)	76.8	14.9	70.2	17.3	63.6	19.9	58.4	22.4	50.1	25.1
CEL (Rules)	**90.0**	**11.9**	**86.0**	**12.1**	**82.3**	**12.6**	**79.2**	**13.1**	**76.8**	**13.7**

The values in bold indicate the best-performing results in every setting. SR is the success rate, and RPL is the relative path length. EL(EA + SI), EL(EA), EL(SI), and EL(STF) represent four common embodied learning (EL) paradigms. EA + SI is a policy learning method that uses the semantic feature cognition (SFC) and the sequential cognitive memory (SCM) to process observations. EA is a policy learning method that uses the SFC to process observations. SI is a transformer-based method that uses the SFC and the SCM to generate semantic tokens. STF is an adaptive embodied learning method based on the student-teacher framework. NBV is a region-based viewpoint optimization computational method.

CEL(Rules) agent demonstrates a 53.3% improvement in SR over the EL(EA + SI) agent. In contrast, in Table 1, in Test Scenario-4 where ten distractors are present, the advantage of the CEL(Rules) agent is only 27.3%. This underscores the advantage of our agent in complex scenarios. Particularly noteworthy, when the occlusion ratio reaches 50%, the NBV and EL(STF) agents exhibit complete test failures in accomplishing the target tracking task. Similarly, when the occlusion ratio increases to 60%, the EL(SI) agent encounters the same issue. In contrast, our agent consistently maintains a high level of SR in the tracking task. Moreover, by introducing different task targets in a completely new, untrained environment, we conduct extension generalization experiments to more systematically evaluate the CEL performance (See Supplementary Note 7 for detailed experiments).

Anomaly active target tracking in real-world scenarios

We deploy the tracker on a UAV platform and conduct tests in two practical outdoor scenarios, Test Scenario-6 and Test Scenario-7, to validate the practical applicability of the CEL framework. The UAV used for the experiments is the Prometheus600 (P600) mesomorphic UAV development platform, equipped with devices such as LiDAR, NX on-board computer, 3-axis optoelectronic pod and RTK. The NX on-board computer model is Jetson Xavier NX, boasting a computing power of 21 TOPS, a 6-core NVIDIA Carmel ARM v8.2 64-bit CPU, NVIDIA Volta architecture for the GPU, 384 NVIDIA CUDA cores + 48 Tensor cores, and a core storage space of 64GB, 8GB DDR4 RAM. The optoelectronic pod model is Q10F, featuring video output at 480P, 30 fps, with a camera performance of 1/3-inch 4MP COMS SENSOR, a camera focal length of 10x optical zoom ($F = 4.9–49$ mm), and a minimum target size of 16×16 pixels. The UAV platform captures images using the optoelectronic pod, predicts movement using the pretrained neural network, and sends movement commands to the controller. The controller outputs desired linear and angular velocities to control the UAV movements.

As depicted in Fig. 6, Test Scenario-6 is designed at an intersection in front of a cafeteria. The scenario includes frequent crowds of people, moving vehicles, dense trees and buildings, all of which result in the prolonged occlusion and intense interference. These common cases significantly impact the UAV performance to detect and track ground targets. Meanwhile, Test Scenario-7 is another challenging target tracking scenario purposefully designed in a playground. In Test Scenario-7, our design incorporates 11 distractors bearing a close resemblance to the task target. To maximize interference with the task target, these distractors and the task target are dressed in identical shirts. Additionally, the task target sporadically utilizes an umbrella to obscure himself, thereby obstructing the UAV visual field in an attempt to evade tracking. These scenarios are tailored to rigorously assess the performance of the CEL framework in highly realistic and demanding conditions.

Figure 6a depicts the UAV configured with the SCM, which integrates multi-view multi-scale historical sequential information to mitigate the presence of distractors and enhance the highlighting of the real task target in Test Scenarios-6 and 7. We visualize the output captured by the UAV optoelectronic pod and present the corresponding typical sequences in the visualization video. From Fig. 6a, it is evident that as the UAV utilizes the sequential memory of the SCM to continuously fuse historical sequential information, several notable observations emerge. Firstly, the cumulative similarity value of the real task target steadily increases over time, while the cumulative similarity of pseudo-targets diminishes correspondingly. Consequently, the peak representing the real task target becomes more pronounced, whereas the peaks of pseudo-targets gradually weaken. Upon closer inspection, it becomes apparent that as the UAV processes observation information through the SCM, it formulates decisions and executes

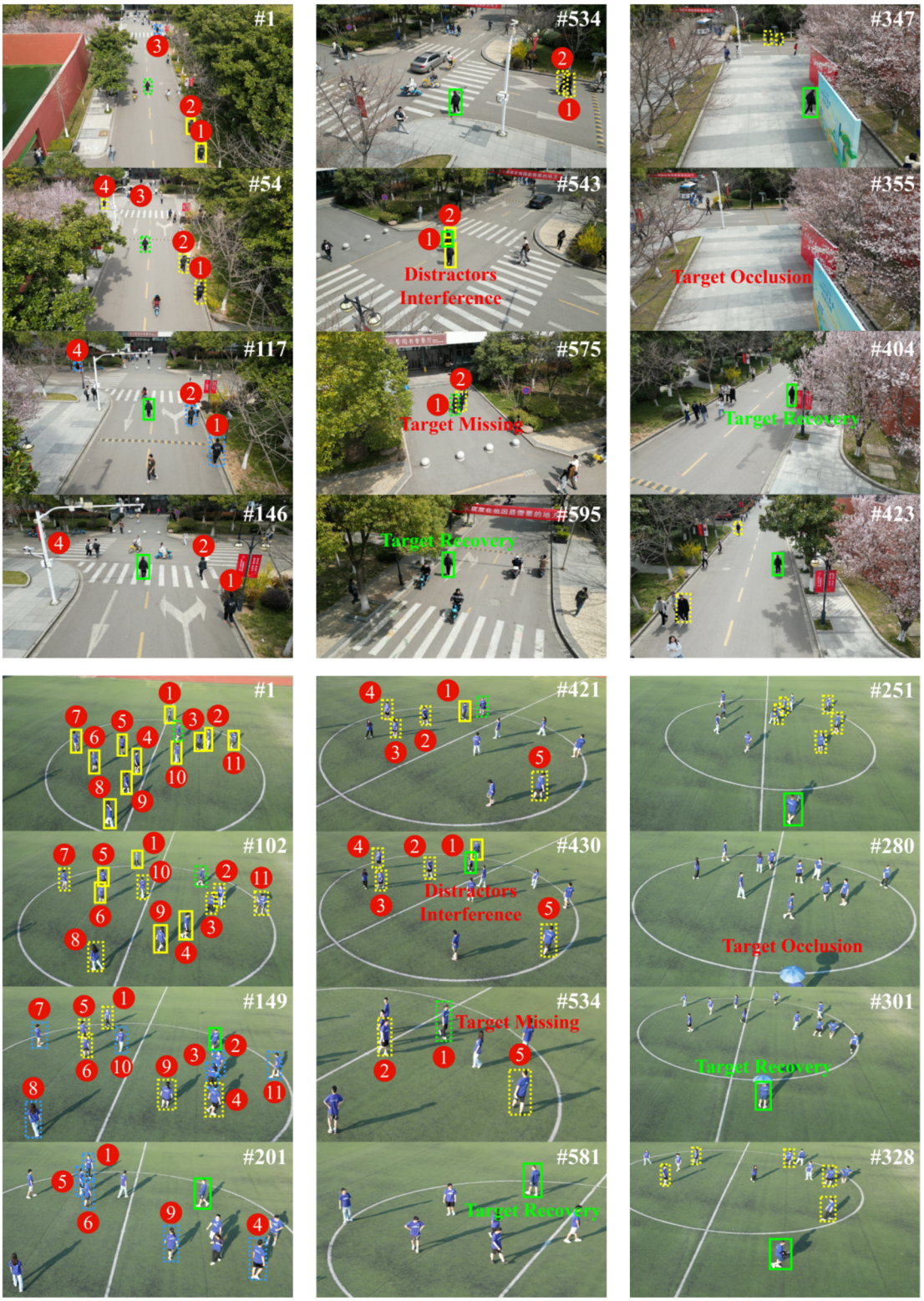

Fig. 6 | Anomaly active target tracking performance in real-world scenarios. a–c illustrate the UAV response to encountering anomalous states during the tracking of the target in Test scenario-6 and Test scenario-7.

maneuvers. As a result, the real task target becomes maximally highlighted, while the number of pseudo-targets is minimized. Moreover, the cumulative similarity values of certain pseudo-targets are sustained at lower levels, and in some cases, these pseudo-targets are completely eliminated. In the depicted figure, various colored boxes denote the cumulative similarity levels of corresponding targets. The green solid box signifies the highest cumulative similarity of the real task target, followed by the green dashed box. Subsequently, the yellow solid box, the yellow dashed box, and the blue dashed box indicate decreasing cumulative similarity values. The presence of the blue dashed box indicates that the similarity of the corresponding

target is maintained at a relatively low level. If the cumulative similarity value continues to decrease, the box disappears entirely. In Test Scenario-6, pseudo-targets are entirely eliminated, leaving only the salient real task target visible. Similarly, in Test Scenario-7, most of the pseudo-targets are eliminated, resulting in the enhanced prominence of the real task target.

Figure 6b, c displays the UAV, equipped with the RRM, applying rules to address bursts of prolonged occlusion and intense interference in Test Scenarios-6 and 7. Figure 6b shows the UAV implementing Rule 1 in response to prolonged occlusion. Test Scenario-6 specifically highlights a situation where the UAV vision is occluded by buildings and further affected by tree shadows. Moreover, Test Scenario-7 delves into a condition where the UAV field of vision is abruptly blocked by an unfurled umbrella carried by the task target. Both scenarios create adverse circumstances, ultimately causing the UAV to deviate and rendering it unable to continue tracking the task target. Nevertheless, the UAV promptly employs Rule 1, effectively resolving prolonged occlusion and successfully resuming its detection and tracking.

Figure 6c illustrates the process of the UAV utilizing Rule 2 to resolve intense interference. In both scenarios, the UAV encounters intense interference from distractors. Particularly in Test Scenario-7, where the distractors closely resemble the appearance of the task target, the interference is the most severe. This leads to erroneous tracking of the distractor and consequently losing sight of the task target. During this process, the matching similarity between the template images and the distractor peaks at a certain scale, viewpoint, and moment, causing the UAV to switch its focus to tracking the distractor instead of the task target. However, as the viewpoint changes, the matching similarity rapidly diminishes. At this point, the UAV detects this anomalous state and promptly employs Rule 2. This prompts the UAV to swiftly return to the nearest landmark to the task target before its loss, utilizing sequential memory to re-establish detection and tracking of the task target. Figure 6c visually depicts all the processes described above. Moreover, we conduct additional qualitative evaluation on a real-world UAV tracking public benchmark to validate the CEL robustness and generalization capability (See Supplementary Note 12 for detailed experiments).

Discussion

In this study, we introduce a novel embodied learning method, namely CEL, addressing the critical limitation of EI in handling severe anomalies of AOT in practical scenarios, such as prolonged occlusion and intense interference. Inspired by the brain's dual decision-making system in cognitive psychology, the CEL demonstrates a significant advancement in the adaptability of AOT systems. The adaptive switching mechanism between the common embodied learning mode and the anomaly handling mode allows for flexible responses to varying environmental conditions. The integration of the ACM, RRM, and AEM collectively enhances the agent's ability to recognize and resolve severe anomalies, reducing the risk of the training divergence and the test failure. This method not only aligns with the principles of EI but also extends it by incorporating categorical perception into the decision-making process. The introduction of the COF offers a systematic approach to defining tracking objectives across different operational states. This innovation effectively addresses the issues of function non-measurability and data confusion in the presence of severe anomalies, facilitating convergence towards feasible solutions and enhancing the overall stability of AOT systems.

The results of our experiments, conducted in both simulated and physical scenarios, demonstrate that the CEL significantly enhances the performance of the UAV in the anomaly active target tracking task. Specifically, the CEL achieves a 361.4% increase in success rate and a 54.4% improvement in task completion efficiency compared to the state-of-the-art methods. The CEL ensures more precise and reliable target tracking in different complex tracking scenarios. These findings underscore the potential of the CEL to revolutionize AOT by effectively mitigating the impact of severe anomalies that traditional methods struggle to handle.

In conclusion, our purposed CEL addresses the critical challenge of maintaining robust and accurate tracking performance in AOT systems when faced with severe physical anomalies like prolonged occlusion and intense interference. These issues are prevalent in practical applications such as autonomous navigation in complex urban environments, search and rescue operations in dynamic disaster zones, security surveillance in cluttered spaces, and wildlife monitoring in natural habitats. As these domains increasingly rely on advanced tracking technologies, the role of CEL becomes progressively more significant. Crucially, the CEL builds upon the recognition of universal anomalous states in tasks, instead of the specific categories of targets and the particular structures of environments. Moreover, the principles underlying the CEL, particularly its adaptive switching mechanism and categorical perception method, hold potential for broader areas of robotics and autonomous systems where adaptability and resilience are crucial. This work paves the way for a new generation of intelligent agents capable of operating effectively in dynamic and unpredictable environments, marking a significant step forward in the development of artificial general intelligence (AGI).

Methods

Predefined expert rules in the RRM

In our purposed CEL, we incorporate two predefined expert rules in the RRM to address the severe anomalies. These expert rules are designed to provide the UAV with robust responses to the specific severe anomalous states, enhancing the resilience of the AOT systems. Specifically, Rule 1 focuses on addressing prolonged occlusion. In this work, prolonged occlusion is defined as the complete obstruction of the task target by obstacles for more than 10 s. When the task target is occluded by obstacles such as buildings and dense trees, the UAV employs Rule 1 to effectively bypass obstacles and re-establish visual contact with the task target on the opposite side of the obstruction. Let $T(x_t, y_t)$ represent the position of the task target, $U(x_u, y_u)$ the current position of the UAV and $O(x_o, y_o)$ be the position of the occluding obstacle. The boundary of the obstacle is defined as

$$B = \{(x, y) | (x - x_o)^2 + (y - y_o)^2 = r^2\}, \tag{4}$$

where r denotes the radius of the obstacle. The UAV then computes a trajectory P that avoids this boundary while maintaining a minimal distance to the task target, given as

$$P = \{(x, y) | (x - x_u)^2 + (y - y_u)^2 = d^2 \wedge (x, y) \notin B\}, \tag{5}$$

where d represents the flight distance. The UAV position is updated according to

$$U(x_u^{new}, y_u^{new}) = \arg \min_{(x,y) \in P} |(x, y) - T(x_t, y_t)|. \tag{6}$$

Moreover, Rule 2 is designed to address intense interference. In our settings, interference is considered to occur when the similarity between the task target and the distractor exceeds 55%. We further categorize the interference level as normal (55–75%), moderate (75–90%), and intense (>90%) based on the degree of similarity. The distractors we deployed typically exhibit a similarity of over 90% with the task target, resulting in the UAV encountering substantial, intense interference during the tracking tasks. When the task target is lost due to intense interference from similar distractors, Rule 2 directs the UAV to return to the most recent location where the task target is last observed, referred to as the landmark. Let $L(x_l, y_l)$ represent the landmark position. The historical sequential observation data, denoted as $\{T(x_t^{(i)}, y_t^{(i)})\}_{i=1}^{n}$, is utilized to guide the UAV back to the landmark via the path, given as

$$P = \{(x, y) | (x - x_u)^2 + (y - y_u)^2 = d^2 \wedge (x, y) = L(x_l, y_l)\}. \tag{7}$$

Upon reaching $L(x_l, y_l)$, the UAV position is updated as

$$U(x_u^{new}, y_u^{new}) = L(x_l, y_l). \tag{8}$$

The UAV leverages the historical sequential observation data to attempt reacquisition of the task target, denoted as $T(x_t^{new}, y_t^{new})$. Rule 2 is crucial in scenarios where the UAV struggles to adapt to intense interference without any explicit guidance.

It is important to emphasize that the CEL does not rely on exhaustively enumerating rules for all possible anomalies. The real world is an open set, and it is neither feasible nor the primary objective of the CEL to create a specific rule for every possible anomaly. The core paradigm of the CEL lies in the adaptive switching between the common embodied learning mode (policy decision-making) and the anomaly handling mode (rule-based reasoning) in response to the normal and anomalous states. Therefore, we select the prolonged occlusion and intense interference as primary anomaly patterns, which have a significant impact on active object tracking tasks. As the primary anomaly patterns are covered, a small number of high-value rules achieves the majority of gains, while additional rules for low-frequency anomalies typically achieve little performance gain (See Supplementary Note 8 for detailed experiments).

The categorical objective function (COF)

In the CEL framework, the agent decision-making process operates in two distinct modes, one relies on the policy network during normal states, while the other leans on the RRM under severe anomalous states. Consequently, the objective function adopts a categorical form, yielding a unified mathematical formulation that systematically describes both measurable normal states and various non-measurable severe anomalous states.

In measurable normal states, the policy network is responsible for optimizing the measurable component of the objective function. Typically, this involves selecting actions to maximize some form of long-term gain, such as cumulative rewards. In this case, the policy network undergoes iterative updates based on the current state and reward signals, enhancing the anomaly active target tracking efficiency. The categorical objective function in measurable normal states is formulated as

$$J_n(s_t, a_t) = \sum_{t=0}^{T} \gamma^t r_t(s_t, a_t), \tag{9}$$

signifies the reward function, γ stands for the discount factor, and T indicates the time horizon of the decision. The objective function is intricately crafted to maximize the long-term cumulative rewards.

In non-measurable severe anomalous states, the agent utilizes the RRM for reasoning to handle the non-measurable component of the objective function. The RRM uses the predefined expert rules to resolve severe anomalies encountered in the anomaly active target tracking task, avoiding complex learning processes. The categorical objective function in non-measurable severe anomalous states is formulated as

$$J_a(s_t, a_t, A_r) = -\eta_a \cdot C(s_t, a_t, A_r) + \gamma_a \cdot R(s_t, a_t, A_r), \tag{10}$$

where A_r denotes the sequence of recovery actions derived from rules-based reasoning. $C(s_t, a_t, A_r)$ represents the recovery cost function. It signifies the cost of executing A_r, which includes the additional interactions consumption, the time delay required for task completion, and the estimation errors in target detection and tracking. $R(s_t, a_t, A_r)$ indicates the recovery reward function, which correlates with the urgency of restoring normal states or mitigating severe anomalies. η_a, γ_a are weighting factors, serving as indicators of the relative importance of rewards and costs in the categorical objective function.

In the CEL framework, the categorical objective function in non-measurable severe anomalous states is tailored to meet the demands of specific scenarios, aiming to swiftly restore normal tracking states when severe anomalies occur, such as intense interference and prolonged occlusion. In such cases, the categorical objective function focuses on rapid anomaly recognition and resolution to mitigate adverse effects on the overall task performance. In the non-measurable states involving intense interference, the categorical objective function prioritizes actions that distinguish between the task target and distractors, redirecting the agent's focus towards the task target, and aiming to minimize interactions with distractors while ensuring consistent tracking of the task target. The categorical objective function in the non-measurable intense interference states is formulated as

$$J_{int}(s_t, a_t, A_r) = -\eta_{int} \cdot C_{int}(s_t, a_t, A_r) + \gamma_{int} \cdot R_{int}(s_t, a_t, A_r), \tag{11}$$

with the specific form of the recovery cost function, given as

$$C_{int}(s_t, a_t, A_r) = \mu_{int}^{(1)} \cdot C_{res}(s_t, a_t) + \mu_{int}^{(2)} \cdot C_t(s_t, A_r) + \mu_{int}^{(3)} \cdot C_e(s_t, a_t), \tag{12}$$

where $C_{res}(s_t, a_t)$ denotes the additional consumption due to interactions, $C_t(s_t, A_r)$ represents the time delay required to task completion, and $C_e(s_t, a_t)$ signifies errors in detection and tracking estimation. $\mu_{int}^{(1)}$, $\mu_{int}^{(2)}$, and $\mu_{int}^{(3)}$ serve as respective weighting factors for individual components of the recovery cost function, delineating their respective contributions to the overall cost and reflecting the relative importance of each component. The specific computational formulations for each are

$$C_{res}(s_t, a_t) = i(s_t, a_t) + l(s_t),\ C_t(s_t, A_r) = t_{rt}(s_t, A_r),\ C_e(s_t, a_t) = e(s_t, a_t), \tag{13}$$

where $i(s_t, a_t)$ denotes the count of erroneous interactions with distractors, indicating instances that distractors are incorrectly tracked, compelling the agent to rely on the rules for reasoning. $l(s_t)$ represents the number of instances of losing track of the task target, given as

$$i(s_t, a_t) = \sum_{k=1}^{N} \mathbf{1}_{\{a_t^{(k)} = A_r\}},\ l(s_t) = \sum_{k=1}^{M} \mathbf{1}_{\{t = t_{los}(s_t)\}}, \tag{14}$$

where $t_{rt}(s_t, A_r)$ denotes the duration from the moment the agent loses track of the task target due to intense interference until it successfully resumes tracking, $t_r(s_t, A_r)$ represents the moment when the task target is rediscovered by the agent using the rules, $t_{los}(s_t)$ signifies the moment when the task target is lost, given as

$$t_{rt}(s_t, A_r) = t_r(s_t, A_r) - t_{los}(s_t), \tag{15}$$

where $e(s_t, a_t)$ denotes the detection and tracking errors, serving as a measure of the agent accuracy in predicting the position of the task target during the detection and tracking process. It is quantified by the disparity between the agent predicted task target position $\widehat{p}$ and the actual position p, given as

$$e(s_t, a_t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (dist(\widehat{p}, p))^2},\ \text{with } dist(\widehat{p}, p) \tag{16}$$

Meanwhile, the specific form of the recovery reward function is given as

$$R_{int}(s_t, a_t, A_r) = R_r + \lambda_{int}^{(1)} \cdot R_{int}^{e}(s_t, A_r) + \lambda_{int}^{(2)} \cdot R_{int}^{s}(s_t, a_t), \tag{17}$$

where $\lambda_{int}^{(1)}$ and $\lambda_{int}^{(2)}$ denote the weighting coefficients, serving as factors for individual components of the recovery reward function. R_r represents the fundamental reward attributed to the successful recovery of detection and tracking, $R_{int}^{e}(s_t, A_r)$ signifies the reward incentivizing the immediate error correction, $R_{int}^{s}(s_t, a_t)$ indicates the reward promoting the system stability.

Their specific formulations are described as

$$R_{int}^{c}(s_t, A_r) = \sum_i \delta(s_i, s_i^{new}, A_r^{(i)}),\ R_{int}^{s}(s_t, a_t) = t_{tra}(s_t) + \bar{e}(s_t, a_t), \quad (18)$$

where $\delta(s_i, s_i^{new}, A_r^{(i)})$ denotes the correction function based on the cumulative similarity value. Here, s_i represents the cumulative similarity value prior to the implementation of A_r, while s_i^{new} signifies the cumulative similarity value following the execution of A_r. If the updated s_i^{new} surpasses the original s_i and exceeds the predetermined threshold θ, the correction function generates a positive value, signifying that the agent successfully rectifies a severe anomaly and potentially resumes tracking of the task target. Conversely, if s_i^{new} does not exhibit a substantial improvement over s_i, the correction function does not generate a reward, given as

$$\delta\left(s_i, s_i^{new}, A_r^{(i)}\right) = \begin{cases} s_i^{new} - s_i, & \text{if} \quad s_i^{new} > s_i + \theta, \\ 0 & \text{otherwise} \end{cases}, \quad (19)$$

where $t_{tra}(s_t)$ denotes the tracking duration and t_{init} represents the starting moment of the tracking. $\bar{e}(s_t, a_t)$ signifies the quality of target detection and tracking, constituting the complement of $e(s_t, a_t)$, thereby reflecting the precision of the tracked target. It exhibits an inverse relationship with $e(s_t, a_t)$, with superior $\bar{e}(s_t, a_t)$ observed as $e(s_t, a_t)$ decreases, given as

$$t_{tra}(s_t) = t_{los}(s_t) - t_{init},\ \bar{e}(s_t, a_t) = 1 - \frac{e(s_t, a_t)}{e_{\max}}. \quad (20)$$

Similarly, the categorical objective function in the non-measurable prolonged occlusion states is

$$J_{ocl}(s_t, a_t, A_r) = -\eta_{ocl} \cdot C_{ocl}(s_t, a_t, A_r) + \gamma_{ocl} \cdot R_{ocl}(s_t, a_t, A_r). \quad (21)$$

The specific forms of the recovery cost function and the recovery reward function are respectively given as

$$C_{ocl}(s_t, a_t, A_r) = \mu_{ocl}^{(1)} \cdot C_{res}(s_t) + \mu_{ocl}^{(2)} \cdot C_t(s_t, A_r) + \mu_{ocl}^{(3)} \cdot C_e(s_t, a_t), \quad (22)$$

$$C_{res}(s_t) = l(s_t),\ C_t(s_t, A_r) = t_{rt}(s_t, A_r),\ C_e(s_t, a_t) = e(s_t, a_t), \quad (23)$$

$$R_{ocl}(s_t, a_t, A_r) = R_r + \lambda_{ocl}^{(1)} \cdot R_{ocl}^{s}(s_t, a_t), \quad (24)$$

$$R_{ocl}^{s}(s_t, a_t) = t_{tra}(s_t) + \bar{e}(s_t, a_t), \quad (25)$$

where $\mu_{ocl}^{(1)}, \mu_{ocl}^{(2)}, \mu_{ocl}^{(3)}$ and $\lambda_{ocl}^{(1)}$ denote the weighting coefficients assigned to individual components of the recovery cost function and the recovery reward function. $C_{res}(s_t)$ encompasses solely $l(s_t)$ and excludes $i(s_t, a_t)$. $R_{ocl}(s_t, a_t, A_r)$ solely incorporates $R_{ocl}^{s}(s_t, a_t)$, without considering R_{ocl}^{c}.

Overall, the behavior of an agent can be conceptualized as optimizing the categorical objective function in measurable normal and non-measurable severe anomalous states. The different categories of objective functions complement each other, collectively guiding the agent's decision-making across various states. In summary, the categorical objective function can be concisely represented as

$$J(s_t, a_t) = \begin{cases} J_n(s_t, a_t) & , \ if \ s_t = s_n, \\ J_a(s_t, a_t, A_r), & if \ s_t = s_a, \end{cases} \quad (26)$$

and $J_a(s_t, a_t, A_r)$ can be further formulated as

$$J_a(s_t, a_t, A_r) = \begin{cases} J_{int}(s_t, a_t, A_r), & if \ s_t = s_{int}, \\ J_{ocl}(s_t, a_t, A_r), & if \ s_t = s_{ocl}, \end{cases} \quad (27)$$

where s_n and s_a denote measurable normal states and non-measurable severe anomalous states. s_{int} and s_{ocl} represent the non-measurable states of intense interference and prolonged occlusion. Assessing the normality of the current state s_t necessitates the utilization of the semantic similarity matrix $\mathbf{M}(s_t)$ (further details of which is elaborated in the subsequent section).

At any given moment, if $\forall j, \exists i,\ \sum_t^{t+n} \mathbf{M}_{xy}^{i}(s_t) \geq \sum_t^{t+n} \mathbf{M}_{xy}^{j}(s_t)$ and $\mathbf{M}(s_t) \neq \mathbf{0}$ are satisfied, it is deduced that s_t coincides with s_n. On the other hand, if at any moment $\forall j, \exists i,\ \sum_t^{t+n} \mathbf{M}_{xy}^{i}(s_t) \geq \sum_t^{t+n} \mathbf{M}_{xy}^{j}(s_t)$ or $\mathbf{M}_{xy}(s_t) = \mathbf{0}$ is met, s_t is deemed to be s_a. Furthermore, if at any moment $\forall j, \exists i,\ \sum_t^{t+n} \mathbf{M}_{xy}^{i}(s_t) \geq \sum_t^{t+n} \mathbf{M}_{xy}^{j}(s_t)$ is fulfilled, s_t is categorized as s_{int}. And when $\mathbf{M}(s_t) = \mathbf{0}$ is attained, s_t is categorized as s_{ocl}.

If the COF is represented in the form of an indicator function, the aforementioned expression can be reformulated as

$$J(s_t, a_t) = \mathbf{1}_{\{n\}} \cdot J_n(s_t, a_t) + \mathbf{1}_{\{ocl\}} \cdot J_{ocl}(s_t, a_t, A_r) + \mathbf{1}_{\{int\}} \cdot J_{inf}(s_t, a_t, A_r), \quad (28)$$

where each indicator function delineates the states experienced by the agent. For instance, $\mathbf{1}_{\{n\}}$ denotes a conditional Boolean expression, assuming a value of 1 when the agent is in measurable normal states and 0 otherwise. Similarly, $\mathbf{1}_{\{ocl\}}$ and $\mathbf{1}_{\{int\}}$ represent the non-measurable states of prolonged occlusion and intense interference, respectively.

When the task target encounters intense interference or prolonged occlusion, the agent recognizes the current non-measurable severe anomalous states. Subsequently, it switches to optimizing $J_a(s_t, a_t, A_r)$, employing the RRM to restore measurable normal states. Once back to the normal state, the agent reverts to utilizing the policy network to optimize $J_n(s_t, a_t)$. This adaptive approach enables the agent to swiftly respond and take actions in non-measurable severe anomalous states while maintaining optimal performance under measurable normal states.

The difference from the standard reward shaping

It is important to distinguish the COF from the standard reward shaping technique in reinforcement learning. While both the COF and the reward shaping aim to improve the agent's performance, they operate at fundamentally different aspects and address distinct challenges. The standard reward shaping typically modifies an original reward function by adding a carefully designed potential-based or heuristic shaping term to provide denser learning feedback ($r^{shaped} = r_t^{env} + F(s_t, a_t, s_{t+1})$), which accelerates the training process and guides the agent towards desired behaviors, addressing the problem of sparse reward. It assumes a consistently static and measurable environment where a single, well-defined reward function can be applied throughout the agent's operation. In contrast, the COF is not a single static reward function. It is a categorical objective that dynamically switches between different objective functions ($J_n, J_{\text{ocl}}, J_{\text{int}}$). This switching is triggered by the cognitive assessment of the current observation state. The primary problem the COF addresses is not the reward sparsity. Instead, it is designed to address the issues of function non-measurability and data confusion caused by severe anomalies, where the standard reward signal becomes meaningless or misleading. The COF is deeply integrated with the CEL framework. When the ACM recognizes an anomaly and the COF switches from the normal objective J_n to an anomalous objective (J_{ocl} or J_{int}), the decision-making mode of the CEL also switches from the learning-based policy network to the RRM, with the objective shifting from maximizing long-term gains to minimizing recovery costs. This cognitive, mode-switching paradigm, inspired by dual-process theory, provides robustness against catastrophic failures that standard reward shaping cannot achieve.

Semantic similarity map

Here, we delve into the computation process of the semantic similarity map. Initially, to extract features from small image segments, we select the APEN and TAdaCNN networks as feature extractors (see Supplementary Note 2 for details). We consider the output of the feature extraction network as a d-dimensional feature vector, where d denotes the dimension of the feature. Let $\mathbf{f}_i \in \mathcal{R}^d$ represent the feature vector of the original image block and

$\mathbf{f}_T \in \mathcal{R}^d$ signify the feature vector of the task target template image. The semantic similarity score, denoted by $S(i, T)$, is computed through cosine similarity, expressed as

$$S(i,T) = \frac{\mathbf{f}_i \cdot \mathbf{f}_T}{\| \mathbf{f}_i \| \| \mathbf{f}_T \|}, \tag{29}$$

where $\cdot$ denotes the dot product of vectors, $\|\mathbf{f}\|$ represents the Euclidean norm of vector $\mathbf{f}$. The cosine similarity operates within the range of $[-1, 1]$, with 1 indicating complete similarity and -1 indicating complete dissimilarity.

The computed semantic similarity scores between every image block and the template image is structured into the semantic similarity matrix $\mathbf{M}(s_t) \in \mathcal{R}^{15\times 15}$, effectively representing the semantic similarity map of the entire raw observation image. Each element $\mathbf{M}_{xy}$ of the matrix corresponds to the similarity score between the small image block (x, y) and the task target template.

In summary, the procedure for computing the semantic similarity map includes the following steps. Firstly, the feature extraction is performed on the image block $\mathbf{I}_i$ and the task target template $\mathbf{T}$ uses the feature extractors. Next, for each image block i and the template T, the cosine similarity score of their respective feature vectors is calculated to obtain the semantic similarity scores $S(i, T)$. Finally, the similarity matrix $\mathbf{M}(s_t)$ is populated based on these computed scores to generate the semantic similarity map.

In the CEL framework, the semantic similarity map serves as a pivotal element, enabling comprehension and analysis of raw observation images at a higher level. It overcomes the constraints posed by the raw pixel data, transforming low-level image information into discernible high-level semantic data. Moreover, owing to its universal applicability, the AOT systems based on the semantic similarity map exhibit efficient adaptability to novel scenarios, demonstrating commendable generalization capabilities.

Evaluation metrics

Here, we provide a detailed description of the two evaluation metrics utilized in our experiments, namely, the success rate (SR) and the relative path length (RPL).

The SR refers to the percentage of successful task completions by the trained agent, ranging from 0 to 1. It is defined as

$$SR = \frac{1}{N_{total}} \sum_{i=1}^{N} N_{success}^{(i)}, \tag{30}$$

where $N_{\text{success}}^{(i)}$ denotes the count of successful trials and N_{total} represents the total number of trials.

The RPL measures the ratio of the actual path length to the shortest feasible path length under the premise of task accomplishment. It is calculated by

$$RPL = \frac{1}{N_{total}} \sum_{i=1}^{N} \frac{L_{act}^{(i)}}{L_{opt}^{(i)}}, \tag{31}$$

where $L_{act}^{(i)}$ denotes the actual path length traversed by the agent to complete the task, $L_{opt}^{(i)}$ represents the shortest feasible path length from the starting point to the endpoint of the task target location. Ideally, the RPL is equal to 1, indicating that the agent takes the shortest path. The RPL greater than 1 indicates that the agent opts for a longer path, which typically signifies a reduction in task efficiency.

By employing these metrics, we rigorously assess the performance of the agent in the anomaly active target tracking task. During the evaluation phase, multiple rounds of trials are typically conducted to mitigate randomness and obtain representative assessment results.

Ethics declarations

The real-world experiments were conducted in public outdoor environments. The primary human subjects tracked as the "task target" in these experiments are the members of the research team who provided full and informed consent to be filmed for the purposes of this study. Other individuals who may appear incidentally in the background of the captured image data were not the subjects of this research. The experiments were designed and conducted in a manner that respects public privacy, and the UAV was operated at a sufficient altitude, and the data processing did not involve the identification or analysis of any individual other than the participating task target. All procedures were performed in accordance with the relevant guidelines and regulations of the Nanjing University of Aeronautics and Astronautics.

Image sources

Figure 6 and Supplementary Fig. 6 contain original images captured during our experiments using the onboard Q10F optoelectronic gimbal pod of the Prometheus600 UAV platform. All image elements and visualizations in these figures are created entirely by the authors. Supplementary Fig. 3 uses images from the publicly available UAV123 benchmark dataset, which is accessible at https://ivul.kaust.edu.sa/benchmark-and-simulator-uav-tracking-dataset under an open-access research license.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The following link provides a detailed practical application demonstration: https://www.youtube.com/playlist?list=PLxto5W315ns7VV1zCS9L8wNKtBowKxUG5.

Code availability

The code used in this study is available on GitHub at https://github.com/DesperateovO/CEL.

Received: 8 January 2025; Accepted: 11 November 2025;
Published online: 27 November 2025

References

1. Almalioglu, Y., Turan, M., Trigoni, N. & Markham, A. Deep learning-based robust positioning for all-weather autonomous driving. *Nat. Mach. Intell.* **4**, 749–760 (2022).
2. Feng, S. et al. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* **615**, 620–627 (2023).
3. Cao, Z. et al. Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning. *Nat. Mach. Intell.* **5**, 145–158 (2023).
4. Zhong, F., Bi, X., Zhang, Y., Zhang, W. & Wang, Y. RSPT: reconstruct surroundings and predict trajectory for generalizable active object tracking. *Proc. AAAI Conf. Artif. Intell.* **37**, 3705–3714 (2023).
5. Bajcsy, A., Loquercio, A., Kumar, A., & Malik, J. Learning vision-based pursuit-evasion robot policies. In *International Conference on Robotics and Automation* (*ICRA*) 9197–9204 (IEEE, 2024).
6. Schedl, D. C., Kurmi, I. & Bimber, O. Search and rescue with airborne optical sectioning. *Nat. Mach. Intell.* **2**, 783–790 (2020).
7. Li, J. et al. Pose-assisted multi-camera collaboration for active object tracking. *Proc. AAAI Conf. Artif. Intell.* **34**, 759–766 (2020).
8. Zhang, Z. et al. Active mechanical haptics with high-fidelity perceptions for immersive virtual reality. *Nat. Mach. Intell.* **5**, 643–655 (2023).
9. Singh, K. P. et al. Ask4help: Learning to leverage an expert for embodied tasks. *Adv. Neural Inf. Process. Syst.* **35**, 16221–16232 (2022).
10. Zhu, Y. et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *International Conference on Robotics and Automation* (*ICRA*) 3357–3364 (IEEE, 2017).

11. Gordon, D. et al. Iqa: visual question answering in interactive environments. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4089–4098 (IEEE, 2018).
12. Li, J. et al. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12123–12132 (IEEE, 2020).
13. Akkaya, I. et al. Solving rubik's cube with a robot hand. Preprint at https://arxiv.org/abs/1910.07113 (2019).
14. Andrychowicz, O. M. et al. Learning dexterous in-hand manipulation. *Int. J. Robot. Res.* **39**, 3–20 (2020).
15. Mitrano, P., McConachie, D. & Berenson, D. Learning where to trust unreliable models in an unstructured world for deformable object manipulation. *Sci. Robot.* **6**, eabd8170 (2021).
16. Luo, W. et al. End-to-end active object tracking via reinforcement learning. In *International Conference on Machine Learning* (*ICML*) 3286–3295 (PMLR, 2018).
17. Zhong, F., Sun, P., Luo, W., Yan, T., & Wang, Y. AD-VAT: An asymmetric dueling mechanism for learning visual active tracking. In *International Conference on Learning Representations* (ICLR, 2019).
18. Zhong, F., Sun, P., Luo, W., Yan, T., & Wang, Y. Towards distraction-robust active visual tracking. In *International Conference on Machine Learning* (*ICML*) 12782–12792 (PMLR, 2021).
19. Lee, Y. et al. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Sci. Robot.* **6**, eabe1315 (2021).
20. Tashakori, A. et al. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nat. Mach. Intell.* **6**, 106–118 (2024).
21. Yuan, D., Chang, X., Huang, P. Y., Liu, Q. & He, Z. Self-supervised deep correlation tracking. *IEEE Trans. Image Process.* **30**, 976–985 (2020).
22. Yuan, D. et al. Active learning for deep visual tracking. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 13284–13296 (2023).
23. Zhong, F., Wu, K., Ci, H., Wang, C. & Chen, H. Empowering embodied visual tracking with visual foundation models and offline RL. In *European Conference on Computer Vision (ECCV)* 139–155 (2024).
24. Li, B., Gan, Z., Chen, D. & Sergey Aleksandrovich, D. UAV maneuvering target tracking in uncertain environments based on deep reinforcement learning and meta-learning. *Remote Sens.* **12**, 3789 (2020).
25. Sadeghi, F., & Levine, S. Cad2rl: real single-image flight without a single real image. Preprint at https://arxiv.org/abs/1611.04201 (2016).
26. Bhagat, S., & Sujit, P. B. UAV target tracking in urban environments using deep reinforcement learning. In *International Conference on Unmanned Aircraft Systems* (*ICUAS*) 694–701 (IEEE, 2020).
27. Jeong, H., Hassani, H., Morari, M., Lee, D. D., & Pappas, G. J. Learning to track dynamic targets in partially known environments. Preprint at https://arxiv.org/abs/2006.10190 (2020).
28. Wilson, M. Six views of embodied cognition. *Psychon. Bull. Rev.* **9**, 625–636 (2002).
29. Anderson, M. L. Embodied cognition: a field guide. *Artif. Intell.* **149**, 91–130 (2003).
30. Howard, D. et al. Evolving embodied intelligence from materials to machines. *Nat. Mach. Intell.* **1**, 12–19 (2019).
31. Gupta, A., Savarese, S., Ganguli, S. & Fei-Fei, L. Embodied intelligence via learning and evolution. *Nat. Commun.* **12**, 5721 (2021).
32. Nygaard, T. F., Martin, C. P., Torresen, J., Glette, K. & Howard, D. Real-world embodied AI through a morphologically adaptive quadruped robot. *Nat. Mach. Intell.* **3**, 410–419 (2021).
33. Cliff, O. M., Saunders, D. L. & Fitch, R. Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle. *Sci. Robot.* **3**, eaat8409 (2018).
34. Haalck, L. et al. Cater: Combined animal tracking & environment reconstruction. *Sci. Adv.* **9**, eadg2094 (2023).
35. Masmitja, I. et al. Dynamic robotic tracking of underwater targets using reinforcement learning. *Sci. Robot.* **8**, eade7811 (2023).
36. Kadambi, A., de Melo, C., Hsieh, C. J., Srivastava, M. & Soatto, S. Incorporating physics into data-driven computer vision. *Nat. Mach. Intell.* **5**, 572–580 (2023).
37. Fang, Z., Jain, A., Sarch, G., Harley, A. W., & Fragkiadaki, K. Move to see better: self-improving embodied object detection. Preprint at https://arxiv.org/abs/2012.00057 (2020).
38. Kotar, K., & Mottaghi, R. Interactron: Embodied adaptive object detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 14860–14869 (IEEE, 2022).
39. Ding, W. et al. Learning to view: Decision transformers for active object detection. In *International Conference on Robotics and Automation* (*ICRA*) 7140–7146 (IEEE, 2023).
40. Kahneman, D. *Thinking Fast and Slow* (Farrar, Straus and Giroux, 2011).
41. Shea, N. et al. Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18**, 186–193 (2014).
42. Taghia, J. et al. Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nat. Commun.* **9**, 2505 (2018).
43. Biswas, D. et al. Mode switching in organisms for solving explore-versus-exploit problems. *Nat. Mach. Intell.* **5**, 1285–1296 (2023).
44. Booch, G. et al. Thinking fast and slow in AI. *Proc. AAAI Conf. Artif. Intell.* **35**, 15042–15046 (2021).
45. Harnad, S. *Categorical Perception* (Nature Publishing Group: Macmillan, 2003).
46. Beer, R. D. The dynamics of active categorical perception in an evolved model agent. *Adapt. Behav.* **11**, 209–243 (2003).
47. Blumberg, H. The measurable boundaries of an arbitrary function. *Acta Math.* **65**, 263–282 (1935).
48. Kharazishvili, A. *Nonmeasurable Sets and Functions* (Elsevier, 2004).
49. Zhu, P. et al. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7380–7399 (2021).
50. Cao, Z., Fu, C., Ye, J., Li, B., & Li, Y. Hift: Hierarchical feature transformer for aerial tracking. In *Proc. of the IEEE/CVF International Conference on Computer Vision* 15457–15466 (IEEE, 2021).
51. Cao, Z., Fu, C., Ye, J., Li, B., & Li, Y. SiamAPN++: siamese attentional aggregation network for real-time UAV tracking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*) 3086–3092 (IEEE, 2021).
52. Ye, J., Fu, C., Zheng, G., Paudel, D. P., & Chen, G. Unsupervised domain adaptation for nighttime aerial tracking. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8896–8905 (IEEE, 2022).
53. Cao, Z. et al. TCTrack: Temporal contexts for aerial tracking. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 14798–14808 (IEEE, 2022)
54. Kim, W. B. & Cho, J. H. Encoding of contextual fear memory in hippocampal-amygdala circuit. *Nat. Commun.* **11**, 1382 (2020).
55. Zaki, Y. et al. Hippocampus and amygdala fear memory engrams re-emerge after contextual fear relapse. *Neuropsychopharmacology* **47**, 1992–2001 (2022).
56. Koenig, N., & Howard, A. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*) 2149–2154 (IEEE, 2004).
57. Shen, L., Huo, C., Xu, N., Han, C. & Wang, Z. Learn how to see: collaborative embodied learning for object detection and camera adjusting. *Proc. AAAI Conf. Artif. Intell.* **38**, 4793–4801 (2024).
58. Devrim Kaba, M., Gokhan Uzunbas, M., & Nam Lim, S. A reinforcement learning approach to the view planning problem. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6933–6941 (IEEE, 2017).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No.BK62427801 and No.62222107.

Author contributions

Q.W. conceived the original idea, provided critical guidance, and supervised the overall project. J.L. designed the framework, conducted the simulation and real-world experiments, analyzed the data, and wrote the original draft of the manuscript. F.Z. contributed to the theoretical development and revised the manuscript. J.J. and H.W. assisted in setting up the UAV platform and conducting the real-world experiments. H.L. participated in the data analysis and validation. K.-K.M. provided senior guidance and helped revise the final manuscript. All authors discussed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-025-00556-6.

Correspondence and requests for materials should be addressed to Fuhui Zhou.

Peer review information *Communications Engineering* thanks Dimitrios Kollias and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: [Miranda Vinay, Rosamund Daw]. A peer review file is available.