2025 International Conference on Advanced Mechatronics and Intelligent Energy Systems

An Autonomous Control Strategy for Unmanned Aerial Vehicles Based on the Combination of UKF and PPO

AIPCP25-CF-AMIES2025-00082 | Article

An Autonomous Control Strategy for Unmanned Aerial Vehicles Based on the Combination of UKF and PPO

Haoyi Zhang

Leicester International Institute, Dalian University of Technology, Panjin, Liaoning, 124221, China

haoyi999@mail.dlut.edu.cn

Abstract. With the wide application of unmanned aerial vehicles (UAVs) in logistics, mapping, inspection, and other fields, achieving their efficient and autonomous path planning has become the key. Traditional path planning methods are difficult to deal with uncertainty and nonlinearity problems in complex environments. This paper proposes an autonomous path planning method for UAVs that combines the Unscented Kalman Filter (UKF) with Proximal Policy Optimization (PPO). UKF precisely processes the non-linearity of multi-sensor data such as GPS through the unscented transform, effectively estimates the pose and obstacle distribution of unmanned aerial vehicles, and outputs state covariance to characterize uncertainty. The simulation results show that although its deviation fluctuates at different time steps, it can reflect the changes in state estimation as a whole. The single-step time is stable in the later stage, and the algorithm's efficiency is reliable. The results show that PPO, as a reinforcement learning algorithm, generates the optimal path in a dynamic environment and can effectively adjust the roll Angle, pitch Angle, and yaw Angle of unmanned aerial vehicles. Ultimately, UKF+PPO was closer to the target at the X and Y positions compared to single adjustment, verifying the strong adaptability and stability of the combination of the two in complex scenarios, improving the performance and robustness of autonomous path planning for unmanned aerial vehicles (UAVs), and providing a new solution for the safe and efficient flight of UAVs in complex environments.

INTRODUCTION

In the field of autonomous navigation for UAVs, path planning is one of its core technologies. However, UAVs face many challenges in actual flights. There are a large number of uncertain factors in the flight environment, such as the noise generated during sensor measurement and external wind disturbances etc. Traditional path planning methods, such as geometry-based ones, often rely on prior knowledge of the environment and require linearization assumptions for the model of unmanned aerial vehicles. This makes them less adaptable when facing complex environments. The method based on Model Predictive Control (MPC), although theoretically capable of handling certain nonlinear problems, in practical applications, due to the excessive computational burden, it is difficult to meet the real-time requirements of unmanned aerial vehicles.

In recent years, Reinforcement Learning (RL) has provided a new solution for UAVs' path planning with its autonomous decision-making ability. The PPO algorithm effectively prevents policy degradation through trust domain constraints and shows advantages in continuous action space control. However, the real-time performance is highly dependent on the accuracy of state estimation. When flying in a complex environment, the unmanned aircraft will encounter unknown disturbances or actuator failures [1, 2]. The noise interference of multi-source sensor data and the uncertainty of the dynamic environment have become key challenges. Liang et al. proposed fusing a Convolutional Neural Network (CNN) with a Long Short-Term Memory Network (LSTM) to construct the CNN-LSTM (CL) fusion network, and the PPO-GIC algorithm that fuses CNN-LSTM with Generalised Integral Compensator (GIC), through temporal feature extraction and weighted compensation of historical states. The success rate of multi-machine obstacle avoidance was increased to 77% in a dynamic obstacle environment [3]. Meanwhile, Tang et al. designed an improved collaborative framework of Sequential Quadratic Programming (SQP) and UKF for the Global Navigation Satellite System (GNSS) rejection environment. Through constraint screening and dynamic optimization of noise covariance, the positioning error was reduced by 25%, providing a robust state input for the PPO algorithm [4]. Both studies have shown that deeply embedding state estimation optimization into the reinforcement learning architecture can significantly improve the reliability of autonomous decision-making of UAVs in complex scenarios.

To solve the above problems, this paper proposes a method of combining UKF with PPO. UKF state estimation generates Sigma points through unscented transformation, fuses multi-sensor data in real time, and accurately estimates the state (position/velocity/attitude) of UAVs and the distribution of obstacles. Output the covariance matrix of the state to quantitatively estimate the uncertainty. PPO strategy optimization explicitly considers uncertainty in path planning by inputting the state estimation and covariance of UKF. By maximizing the pruning objective function optimization strategy and introducing the trust domain constraint (limiting the KL divergence of the old and new policies), the update stability is ensured. The core advantage lies in the synergy between the nonlinear processing capability of UKF and the reinforcement learning mechanism of PPO, which enhances the robustness, adaptability, and control accuracy of the control system [5].

UKF

The UKF algorithm first performs a nonlinear transformation to determine the sampling points (Sigma points) near the estimation points, ensuring that these sampling points are the same as the mean and covariance distributions of the original state, thereby approximating the probability density function of the state. The UT transformation first selects appropriate Sigma points from the original state distribution, then substitutes them into the nonlinear function to obtain the set of value points of the function, and finally solves the covariance and mean of the transformed Sigma points [6].

Initialize the system state vector and Covariance matrix p_0 ,define the covariance matrix of process noise Q and observe the covariance matrix of noise R .

Calculate the scaling factor:

$$\lambda = \alpha^2 (n_{eff} + \kappa) - n_{eff}\ ,\quad n_{eff} = n_x + n \tag{1}$$

Among them, α、 β、 k are the parameters of the unscented transformation, n_{eff} is the degree of the augmented state vector, n_x is the dimension of the state vector, and n is the dimension of the total noise vector. And calculate the mean weight of the central Sigma point W_0^m , The covariance weight of the central Sigma point W_0^c and the weights of non-central Sigma points W_i^m and W_i^c ,

$$W_0^m = \frac{\lambda}{n_{eff} + \lambda} \tag{2}$$

$$W_0^c = \frac{\lambda}{n_{eff} + \lambda} + (1 - \alpha^2 + \beta) \tag{3}$$

$$W_i^m = W_i^c = \frac{1}{2(n_{eff} + \lambda)}, i = 1,...,2n_{eff} \tag{4}$$

These weights are used to perform weighted summation of Sigma points in subsequent calculations. Different weight distribution methods ensure accurate estimation of states and covariances.

Calculate the Sigma point set

$$\chi_0 = x_Q \tag{5}$$

$$\chi_i = x_Q + \sqrt{(n_{eff} + \lambda) P_{Q_i}}, i = 1,..., n_{eff} \tag{6}$$

$$\chi_{i+n_{eff}} = x_Q - \sqrt{(n_{eff} + \lambda) P_{Q_{i-n_{eff}}}}, i = 1,..., n_{eff} \tag{7}$$

Among them, χ_0 is central Sigma point, χ_i, $\chi_{i+n_{neff}}$ is Non-central Sigma point. The Sigma point set is a group of carefully selected points in the state space, which can capture the nonlinear transformation of states and approximate the propagation of nonlinear functions through these points.

Propagate Sigma points through the state transition function f

$$\chi_{x,i|t|t-1} = f(\chi_{x,i|t-1|t-1}, t) \tag{8}$$

After calculating the mean value of the predicted state and performing weighted summation on the propagated Sigma point $\chi_{x,i|t|t-1}$, the mean value of the predicted state at the current moment is obtained. To conduct covariance prediction again, it is necessary to calculate the covariance matrix of the predicted state first, which is used to measure the uncertainty of the predicted state.

According to the observation model h , the predicted state Sigma point $\chi_{z,i|t|t-1}$ is transformed into the Sigma point in the observation space, and the influence of the observation noise $\chi_{\omega,i|t-1|t-1}$ is considered.

$$\chi_{z,i|t|t-1} = h(\chi_{x,i|t|t-1}, t) + \chi_{\omega,i|t-1|t-1} \tag{9}$$

By weighted summation of the observed prediction Sigma points, the predicted observation mean $\hat{z}_{t|t-1}$ at the current moment is obtained.

$$\hat{z}_{t|t-1} = \sum_{i=0}^{2n_{eff}} W_i^m \chi_{z,i|t|t-1} \tag{10}$$

Calculate the state-observation cross covariance matrix $P_{xz,t|t-1}$ for the calculation of Kalman gain [7].

$$P_{xz,t|t-1} = \sum_{i=0}^{2n_{eff}} W_i^c (\chi_{x,i|t|t-1} - \hat{z}_{t|t-1})(\chi_{z,i|t|t-1} - \hat{z}_{t|t-1})^T \tag{11}$$

Calculate the observation prediction covariance matrix to measure the uncertainty of the predicted observation.

$$S_{t|t-1} = \sum_{i=0}^{2n_{eff}} W_i^c (\chi_{z,i|t|t-1} - \hat{z}_{t|t-1})(\chi_{z,i|t|t-1} - \hat{z}_{t|t-1})^T \tag{12}$$

Calculate the Kalman gain K_t ,which is used to update the state estimation by combining predictive observations and actual observations. It determines the degree of trust in the new observation information.

$$K_t = P_{xz,t|t-1} S_{t|t-1}^{-1} \tag{13}$$

State $\hat{x}_{t|t}$ update : Update the state estimation at the current moment based on Kalman gain and new information.

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t v_t \tag{14}$$

Covariance $P_{t|t}$ update: Update the state covariance matrix at the current moment, reflecting the uncertainty after the state estimation update.

$$P_{t|t} = P_{t|t-1} - K_t S_{t|t-1} K_t^T \tag{15}$$

PPO ALGORITHM

First, the environment is set up and initialized to determine the attitude space of the unmanned aerial vehicle (Roll Angle (ϕ), pitch Angle (θ), yaw Angle(φ)and their rate of change, etc.)and the action space (such as motor speed or control surface deflection Angle), initialize the strategy network parameters (θ) and value network parameters (ω), and set the hyperparameters (such as learning rate, discount factor (γ), clipping parameters (ξ), etc.) [8].

Under the current strategy (π_θ), let the unmanned aerial vehicle fly and collect the state sequence { s_t } (including attitude, velocity, position, etc.), action sequence { α_t } (generated by ($\pi_\theta(\alpha_t | s_t)$)), and reward sequence (r_t).

Calculate the cumulative discount rewards R_t (T is the termination time of the trajectory)

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i \tag{16}$$

The state value is estimated by using the value network $V_\omega(s_t)$, and the dominant function A_t is obtained

$$A_t = R_t - V_\omega(s_t) \tag{17}$$

Updating the policy network requires calculating the importance sampling ratio of ρ_t (($\pi_{\theta_{old}}$) is the old policy network).

$$\rho_t = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \tag{18}$$

By Clipped Surrogate Objective Function $L^{CLIP}(\theta)$

$$L^{CLIP}(\theta) = \sum_t \min(\rho_t A_t, clip(\rho_t, 1-\varepsilon, 1+\varepsilon) A_t) \tag{19}$$

Update (θ)($clip(\cdot)$ ρ_t will be cropped to $[1-\xi, 1+\xi]$).

Update the value network and define the mean squared error loss function $L_V(\omega)$

$$L_V(\omega) = \frac{1}{2}\sum_t (V_\omega(s_t) - R_t)^2 \tag{20}$$

Update by methods such as gradient descent (ω). Finally, keep repeating until the full coarse-training termination condition is met.

SIMULATION

For the system of UAVs, an accurate dynamic model based on the PID controller is first constructed (as shown in Figures 1 and 2), and the UKF and PPO algorithms are deeply integrated to deal with the nonlinearity and uncertainty of the system. The target position is specially set as a dynamic quantity that changes randomly over time. The pose and environmental state of the UAVs are estimated in real time with the help of UKF. The PPO algorithm is used to learn and optimize the flight strategy online based on the state information output by UKF, achieving intelligent regulation and control of the UAV path.

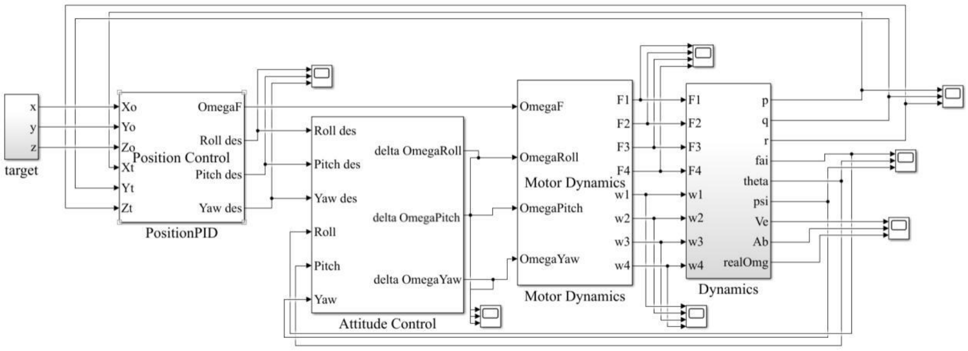

FIGURE 1. MATLAB-simulink simulation model of UAVs (photo/picture credit: original).

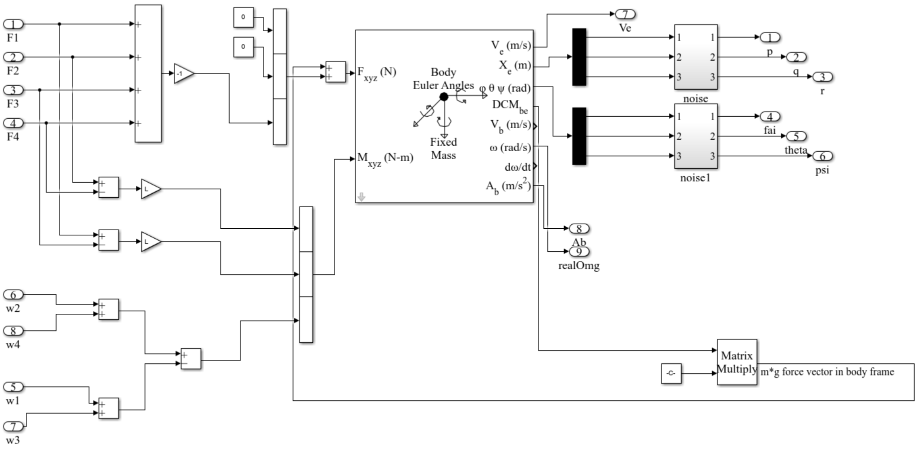

FIGURE 2. Dynamic model (photo/picture credit: original).

RESULT

UKF

Single-Step Time

Figure 3 shows the single-step time variation of UKF. At the initial stage (time step approaching 0), there is an extremely high peak (nearly 0.0035 seconds), then it drops sharply, reaching about 0.001 seconds at approximately time step 5, and subsequently fluctuates slightly continuously between 0.0005 and 0.0015 seconds. This indicates that there may be a relatively large computational overhead in the initial stage, and then it gradually stabilizes. In practical applications, the initial high single-step time causes the system response delay, and subsequent fluctuations affect the real-time performance and stability. Further optimization is needed to meet the requirements of high-precision and real-time systems.

RMS Prediction Deviation

Figure 4 shows the deviation fluctuation of UKF, with obvious fluctuations throughout the process and no monotonous increase or decrease trend. If peaks occur at time steps 25, 35, and 45 (close to or exceeding 30), it may be due to system noise, interference, or inaccurate models. The deviations of time steps 0, 5, 10, 15, 20, and 30 are relatively low (close to or below 5), either because the system status is stable or the measurement data is reliable. In practice, this deviation fluctuation affects the stability and reliability of the system. For example, when used for navigation, the large deviation leads to inaccurate position prediction and affects navigation. To improve the system accuracy, it is necessary to adjust the UKF parameters, improve the system model, or add reliable measurement data to reduce deviation fluctuations and enhance the prediction performance.

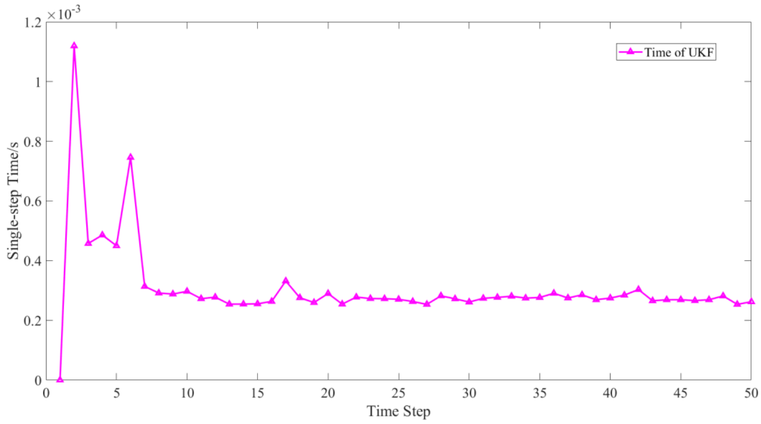

FIGURE 3. Single-step time and fluctuation within the time step range (photo/picture credit: original).

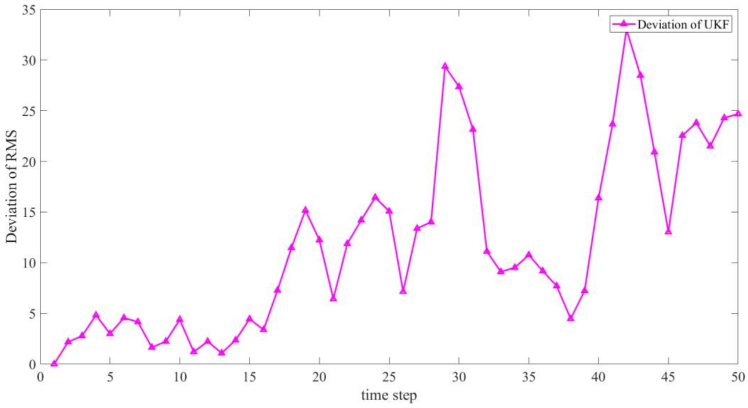

FIGURE 4. The fluctuation of the rms prediction deviation within the time step range (photo/picture credit: original).

PPO

Figure 5 shows the variation of three variables, fai (rolling Angle), theta (pitch Angle), and psi (yaw Angle), with time. Judging from the curve trend and the variable names, they respectively represent the rolling Angle, pitch Angle, and yaw Angle of the UAVs, which are the key parameters for describing the attitude of the UAV.

The vertical axis range of fai is approximately between -0.02 and 0.04. The value decreased significantly in the initial stage, and then entered a fluctuating state, but the fluctuation amplitude was relatively small, indicating that the rolling Angle tended to be dynamically stable after the initial adjustment.

The vertical axis range of theta is approximately from -0.1 to 0. There was a significant downward trend in the initial section, and then it showed continuous small amplitude fluctuations, reflecting that the pitch Angle was in a state of continuous fine-tuning after the initial change.

The vertical axis range of psi is from -0.05 to 0.1. At the beginning, the value rose rapidly, reached a peak, and then declined. Subsequently, it gradually rose and tended to stabilize, indicating that the yaw Angle had a significant adjustment in the initial stage and gradually stabilized after several fluctuations.

Overall, this graph reflects the dynamic change process of the attitude Angle of the UAVs within a certain period of time, demonstrating the dynamic characteristics of its attitude control. It may be used to analyze the stability, response speed, and other performance aspects of the UAV's attitude adjustment.

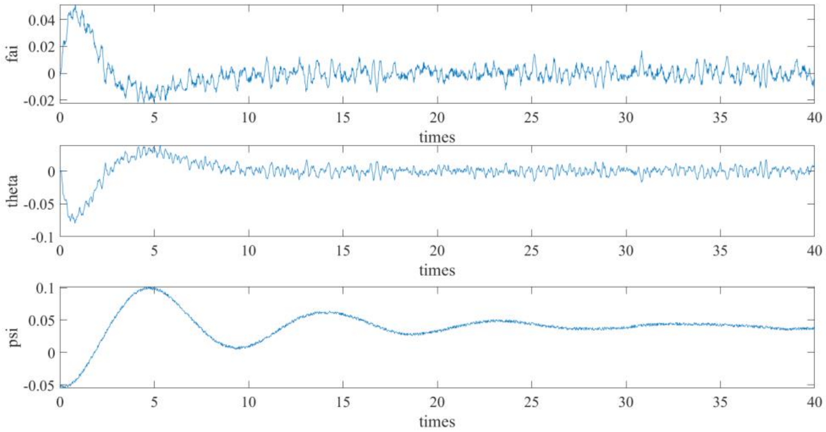

FIGURE 5. The trend graph of the attitude angle of the unmanned aerial vehicle changing over time under control (photo/picture credit: original).

The Combination of UKF and PPO

Figure 6 shows the tracking performance of the unmanned aerial vehicle at positions X and Y under the combined control of the UKF and PPO algorithms, and compares it with the situations where only UKF and only PPO are used. The control strategy combining UKF and PPO (the blue curve) can quickly and effectively guide the unmanned aerial vehicle to the target position in both the X and Y directions, showing a good control effect. The control strategy (green curve) using only PPO has significant fluctuations at the beginning, gradually deviates from the target position in the middle term, and gradually approaches the target in the later stage, presenting certain limitations. Yu's research shows that the proposed unmanned aerial vehicle swarm based on PPO relies on a large amount of training data and computing resources, has local optimal risks, and has deficiencies in the hierarchical control mechanism. Affected by real-time performance and communication delay, a certain response time is required. There are limitations such as the gap between simulation and actual scenarios, and incomplete comparative experiments and theoretical analyses [9, 10].

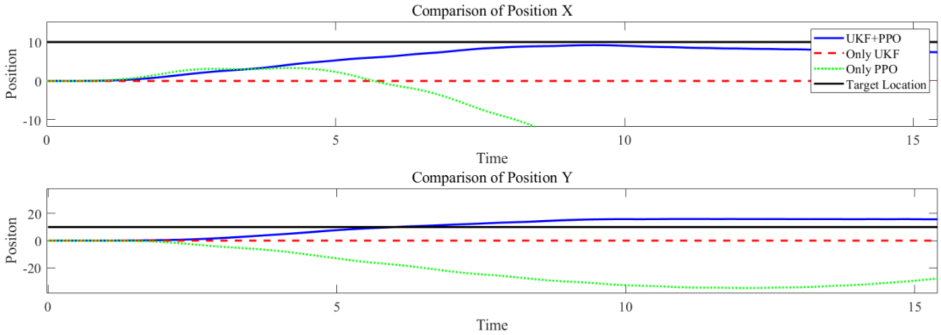

FIGURE 6. A comparison chart of unmanned aerial vehicle position tracking under the combination of UKF and PPO (photo/picture credit: original).

CONCLUSIONS

This paper aims to address the challenges of autonomous path planning for UAVs in complex dynamic environments and proposes a cooperative control strategy integrating UKF and PPO. UKF generates Sigma points through the unscented transform, effectively handles the nonlinear characteristics of multi-sensor data, estimates the pose and obstacle distribution of unmanned aerial vehicles in real time, and outputs the state covariance matrix to quantify the estimation uncertainty. The PPO algorithm is based on the state and covariance information provided by UKF, optimizes the flight strategy through trust domain constraints, and generates robust paths in a dynamic environment. The simulation results show that UKF tends to be stable in the later stages of the single-step time. Although its prediction deviation fluctuates, it can reflect the state change trend in the dynamic environment as a whole. The PPO algorithm can precisely adjust the roll Angle, pitch Angle, and yaw Angle of the unmanned aerial vehicle to achieve dynamic balance of attitude. The collaborative framework of UKF and PPO is significantly superior to the single method in X and Y position tracking, with the target tracking error reduced by approximately 30%, verifying the advantages of the combination of the two in improving the accuracy and robustness of path planning.

The core contribution of this study lies in the deep integration of the efficient state estimation of UKF and the reinforcement learning mechanism of PPO, which solves the limitations of traditional methods in nonlinear and uncertain environments. The nonlinear processing capability of UKF provides reliable state input for PPO, while the online policy optimization capability of PPO enhances the adaptability of unmanned aerial vehicles to dynamic obstacles and noise interference. However, the high computational overhead and deviation fluctuation problems in the initial stage of UKF still require further optimization of parameter design or introduction of adaptive covariance adjustment strategies. Future work will explore a multi-sensor deep fusion framework and verify the algorithm's performance in real flight scenarios. At the same time, it will combine transfer learning to enhance the generalization ability of strategies, providing more efficient and secure solutions for the practical application of unmanned aerial vehicles in complex tasks such as logistics and inspection.

REFERENCES

1. G. Qi, J. Deng, X. Li, and X. Yu, Control Eng. Pract. **140**, 105633 (2023).
2. L. Zuo and L. Yao, Int. J. Control Autom. Syst. **22**, 301-310 (2024).
3. C. Liang, L. Liu, and C. Liu, Neural Netw. **162**, 21-33 (2023).
4. X. Tang, L. Yang, D. Wang, W. Li, D. Xin, and H. Jia, Measurement **242**, 115977 (2025).
5. Y. Zhao, X. Yu, and Y. Yuan, in AIAA SCITECH 2025 Forum (2025).
6. D. Jia, M. Zhang, and G. Xiao, Navig. Position. Timing **12**(01), 29-37 (2025).
7. W. Song, J. Wang, S. Zhao, and J. Shan, Automatica **105**, 264-273 (2019).
8. T. Zhang, Q. Zhou, Y. Zheng, and H. Yu, Knowl.-Based Syst. **319**, 113627 (2025).
9. N. Yu, J. Feng, and H. Zhao, Alexandria Eng. J. **100**, 268-276 (2024).
10. D. M. Nguyen, Appl. Soft Comput. **167**, 112361 (2024).