

Article

Distributed Cooperative Self-Localization Algorithm for Multi-UAVs in Aerial Gaming Scenarios

Qing Liang ^{1,2}, Yingzhi Ouyang ¹ and Hui Li ^{1,*}

¹ School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 610054, China; 202312100627@std.uestc.edu.cn (Q.L.); 202221100206@std.uestc.edu.cn (Y.O.)

² Avic Chengdu Aircraft Design & Research Institute, Chengdu 610091, China

* Correspondence: kelly.li@uestc.edu.cn

Abstract

Accurate and consistent self-localization is essential for multi-UAV aerial missions in complex dynamic environments. However, communication constraints and heterogeneous sensor reliability variations often lead to cumulative localization errors and degraded robustness in conventional fusion frameworks. To address these challenges, this paper proposes a distributed cooperative localization framework integrating deep temporal feature learning, heterogeneous multi-sensor fusion, and consistency-aware distributed state estimation. First, an LSTM-based staged fusion strategy is designed to integrate VIO, GPS, and UWB measurements for accurate single-UAV localization. Second, a Squeeze-and-Excitation LSTM Self-Attention (SE-LSTM-SA) network is developed to adaptively recalibrate heterogeneous sensor channels and enhance temporal feature extraction under dynamic sensing conditions. Finally, a consistency-aware distributed fusion mechanism based on the Labeled Multi-Bernoulli (LMB) framework is introduced to improve inter-UAV state consistency through iterative local-neighbor information exchange. Experiments conducted on the XTDrone platform demonstrate that the proposed framework achieves superior localization accuracy compared with traditional EKF and conventional LSTM-based methods. Specifically, the proposed method achieves lower RMSE, MAE, and Maximum Prediction Error (MaxPE), while significantly improving global consistency performance. Experimental results demonstrate that the proposed framework provides accurate and consistent localization performance for multi-UAV systems in complex dynamic environments.

Keywords: co-localization; LSTM network; coherent localization



Academic Editor: Konstantinos Kontis

Received: 2 April 2026

Revised: 31 May 2026

Accepted: 22 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

UAVs have become an essential platform in multi-agent aerial mission scenarios by virtue of their high mobility, low cost, and reduced human operational risk. Not only can they be deployed quickly, adjusted flexibly, and transferred on demand in changing environments, but they also play a key role in complex application tasks such as multi-target tracking, information reconnaissance, and mission support.

In the process of multi-UAV cooperative operations, communication resources are a critical constraint. In some cases, there may be a variety of interfering factors, such as complex electromagnetic interference applied by adversaries. In a communication-restricted environment, the communication between UAVs and ground terminals or within the swarm will be heavily interfered with, and reliable large-scale, low-latency communication cannot

be guaranteed [1]. Therefore, UAV clusters must possess the ability to autonomously complete tasks under conditions of no external command and global communication network obstruction. This demand has strongly motivated the in-depth research and application exploration of distributed perception fusion and autonomous positioning technology [2].

Traditional stand-alone systems face many challenges in complex environments. Single sensors are limited by data accuracy, noise interference, and environmental adaptability, and cannot meet the perception requirements in complex dynamic environments. For example, GPS may fail under signal occlusion or external interference; vision sensors are unstable in environments with drastic changes in illumination; and IMUs have cumulative errors over long periods of operation [3,4]. In order to improve the sensing accuracy and robustness, and to extract environmental information more effectively, multi-source sensor fusion has become a mainstream solution [5]. Fusion of multi-sensor data can enhance the system's anti-interference ability and overall robustness, and maintain good state estimation when some sensors fail through redundant calibration.

However, the single-machine multi-sensor system still has its limitations. The installation position of sensors and platform motion constrain the observation viewpoint and coverage, and global knowledge cannot be obtained. Furthermore, the single-aircraft platform has limitations in energy, computing resources, and communication bandwidth, making it difficult to realize the deep fusion of multidimensional information. These shortcomings have prompted researchers to explore the fusion localization method under multi-UAS, to extend the observation range through multi-platform cooperation, to further improve the robustness and accuracy of the system through distributed sensing and data sharing, and to provide more reliable localization and sensing support for mission accomplishment in complex dynamic environments [6].

Multi-UAS fusion of relative positioning and global position information maintains the stability of the overall state estimation [7]. For example, combining global position information (e.g., GPS) with local relative position information (e.g., UWB and visual odometry) enables globally consistent position estimation and error compensation with the help of fusion strategies such as the Extended Kalman Filter (EKF) or neural networks. Local short-term position changes are acquired by high-frequency sensors (e.g., IMUs and visual sensors), while global long-duration drifts are corrected by global observation information (e.g., GPS and co-observation). This fusion framework compensates for the limitations of single observation, improves resource utilization, sensing range, and anti-interference capability.

Sensor fusion methods commonly used in multi-UAS can be classified into centralized fusion, decentralized fusion, and distributed fusion [8]. Traditional centralized fusion methods rely on a central node to collect and process sensor data from all UAVs. As a result, the central node is prone to system performance bottlenecks in complex mission environments with limited communication and incomplete information. This leads to the difficulty of centralized fusion methods to meet the demand for fast and reliable cooperative sensing and localization. Therefore, distributed fusion has become an emerging paradigm in recent years [9].

The distributed fusion method realizes data complementarity and error correction by regularly sharing state estimation and local observation information among UAVs. It can effectively reduce the influence of noise in individual sensor data and realize more accurate state estimation. Especially in communication-constrained environments, timely state updates can be achieved through low-latency data transmission between neighboring aircraft, ensuring mission continuity and accuracy [10]. Recent advancements in distributed control theory [11,12] have provided solid theoretical foundations for multi-agent cooperative coordination. Built upon these concepts, distributed cooperative localization enables UAV

clusters to maintain stable operation of the overall system in the face of local sensor failure or interference. Thus, it provides solid technical support for autonomous decision-making and cooperative operation in high-risk mission scenarios.

However, existing distributed localization methods still face significant challenges in communication-constrained and dynamic multi-UAV scenarios. In such environments, severe communication constraints cause traditional filtering methods to suffer from rapid error accumulation. Moreover, while data-driven methods have been introduced for sensor fusion, they often struggle to dynamically assign weights to heterogeneous sensors under noise, and existing distributed algorithms lack an efficient mechanism to promote global consistency among UAVs without overwhelming the communication bandwidth. Therefore, existing works fall short in addressing the coupled challenge of heterogeneous sensor fusion and global state consistency under restricted communication.

To bridge this specific gap, this paper focuses on multi-sensor fusion, distributed collaborative sensing, and autonomous localization for multi-UAV systems. Existing LSTM-based temporal fusion models generally treat heterogeneous sensor channels equally during feature extraction, making them less effective in handling unequal feature contributions and noise corruption under dynamic environments. To address this limitation, a multi-source fusion cooperative localization algorithm based on a compression-excited LSTM-SA temporal network and distributed coherent labeled multi-Bernoulli (LMB) filtering is proposed. By embedding the SE mechanism, the framework enables adaptive channel-wise recalibration according to learned feature importance, thereby improving the robustness of heterogeneous sensor feature representations under varying noise conditions. Through phased information fusion, temporal feature optimization, and distributed consistency data processing, the proposed algorithm improves localization accuracy while maintaining computational feasibility for distributed deployment.

The main contributions of this study are summarized as follows:

- **Staged Multi-Sensor Fusion via Deep Learning:** An LSTM-based staged fusion strategy is designed to mitigate the limitations of fixed-weight filters. It effectively integrates heterogeneous data (VIO, GPS, UWB) for accurate single-UAV localization, capturing complex temporal dependencies under dynamic conditions.
- **Consistency-Aware Distributed Weight Update Mechanism:** A Local Consistency Weight Update (LCWU) mechanism based on the distributed consistency LMB algorithm is introduced. This addresses the problem of positioning inconsistency in information-blocking environments, enabling UAVs to dynamically adjust relative weights and promote global consistency through iterative local-neighbor information fusion.
- **Heterogeneous Channel-Calibrated SE-LSTM-SA Network:** A compression-excited LSTM-SA network is developed to address the unequal feature contribution problem of heterogeneous sensor channels under noisy environments. Through adaptive channel-wise recalibration, the network dynamically reweights heterogeneous sensing features according to learned feature importance, improving the robustness and adaptive feature representation capability of the localization framework.

2. Related Works

High-precision and reliable localization is a fundamental prerequisite for UAV swarm operations in dynamic environments [13]. Compared with autonomous localization, cooperative localization significantly improves system observability and robustness by exploiting inter-UAV geometric constraints and shared sensing information. Traditional state-estimation frameworks, such as Extended Kalman Filter (EKF)-based fusion, factor-graph optimization [14], and distributed consensus filtering, remain widely adopted baselines for

this task. While these filtering and optimization-based methods provide rigorous mathematical guarantees under standard conditions, they often rely on accurate predefined dynamic models and can suffer from linearization errors or high computational overhead in highly dynamic multi-UAV environments.

To operate in GNSS-denied environments, recent studies have explored various distributed cooperative strategies. Geometric-based methods, such as terrain elevation matching and relative triangulation [15,16], have been proposed to correct inertial navigation drifts. However, they are highly sensitive to initial configurations and high-speed flight dispersion. To address communication uncertainties, distributed information transfer techniques like the Belief Propagation (BP) algorithm [17] and covariance intersection methods have been introduced. Although BP-based methods improve robustness under packet loss, they still rely heavily on handcrafted filtering structures and lack the adaptive capacity to perform adaptive heterogeneous feature integration under unpredictable noise.

Multi-source sensor fusion (e.g., VIO, UWB, LIDAR) is critical to overcoming the limitations of single sensors [18–21]. To improve adaptive capabilities, deep learning techniques, particularly Long Short-Term Memory (LSTM) networks [22] and their attention-enhanced variants [23–26], have been increasingly adopted for time-series sensor fusion. These data-driven models successfully capture complex temporal relationships that traditional EKF methods struggle with. However, existing LSTM-based temporal fusion models generally treat heterogeneous sensor channels equally during feature extraction. This equal-weighting characteristic limits their ability to adaptively handle unequal feature contributions.

Despite recent advances in distributed localization and deep learning-based sensor fusion, existing methods still face two key limitations under communication-constrained multi-UAV environments: (1) insufficient robustness to heterogeneous sensor reliability variations under dynamic noise; and (2) a lack of effective distributed consistency maintenance mechanisms when global communication is severely restricted.

These limitations directly motivate the distributed cooperative localization framework proposed in this work.

3. Multi-UAV Distributed Coherent Self-Localization Algorithm Based on SE-LSTM-SA

3.1. LSTM-Based Perception Fusion Self-Localization Algorithm for a Single UAV

Traditional multisensor fusion methods mainly rely on filters or optimization frameworks. These methods have certain advantages in specific scenarios, but when dealing with nonlinear and multimodal data in complex environments, it is often difficult to cope with their dynamic changes. For the localization of a single UAV, a multi-stage sense-fusion self-localization method based on a long short-term memory (LSTM) network is designed. The method integrates data from heterogeneous sensors and solves the problem of processing high-frequency dynamic data from sensors such as IMUs and cameras. And it provides global correction by GPS and UWB sensors to improve the positioning accuracy of a single UAV system equipped with multiple sensors.

In this paper, we design a multi-source sensor fusion localization algorithm for UAV based on LSTM network. The multi-source sensor observations obtained at each moment are regarded as a time series, and the data from four types of sensors, namely IMU, GPS, UWB, and camera, are fused through the steps of data preprocessing, network design, and model training. The final realization is to output the position information of the UAV. The fusion process is shown in Figure 1.

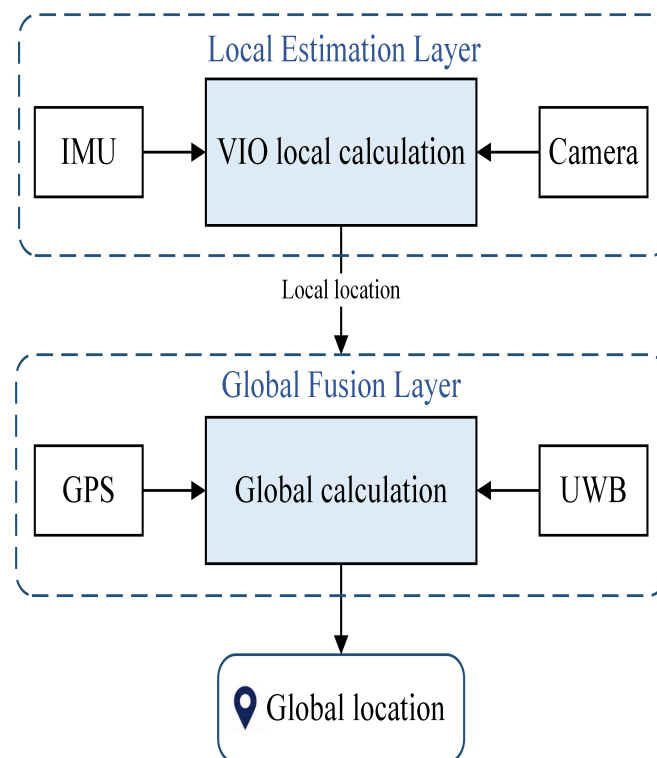


Figure 1. Staged integration process.

3.1.1. Timing Data Input

In the real-time input layer of the simulation data, the four sensor data streams are read from different channels of the ROS communication node. In the simulation platform, the raw data from IMU, GPS, UWB, and camera are unified under the same coordinate system through spatial and temporal alignment to ensure the time synchronization and spatial consistency of the data.

IMU data are characterized by continuous changes and sensitive to time series changes. In the data processing stage, IMU data is a time series containing acceleration (a_x, a_y, a_z) and angular velocity (w_x, w_y, w_z) . The data are fused with binocular vision sensor data as intermediate states in the model, together with the error correction model, to realize real-time tracking and prediction of the flight position of the moving body.

In the overall process, the visual sensor (binocular camera) extracts environmental information, target detection and position estimation from the image data. The processed binocular camera data and IMU data are input into the VIO [27] local fusion localization module, and the EKF local fusion localization method is used for camera position estimation. The system first performs feature extraction and matching on the image at the front-end to generate the key frame sequence. Meanwhile, the high-frequency data from IMU is utilized for pre-integration to capture the dynamic motion information of the platform. Subsequently, the visual observation constraints are fused with the IMU pre-integration results based on the sliding window EKF filtering method to solve the problem, so as to estimate the camera position, velocity, and sensor bias. In this case, the UWB distance measurements and the GPS localization data are independently input to the LSTM network from two channels as global correction quantities.

3.1.2. Localized Integration

Multi-computer multi-source sensor fusion localization is divided into three phases, which are local fusion phase, global correction phase, and distributed sensing sharing, and its hierarchical processing is achieved through multi-level local and global computation.

At the single-machine level, the fusion phase contains two processes, local computation and global computation.

First, the local computation is done independently by each UAV, which is mainly responsible for processing the binocular camera and IMU high-frequency dynamic data to obtain the local VIO results, which can provide short-term motion estimation for the global correction. For local fusion, the Extended Kalman Filter (EKF) is used, and literature [28,29] provides a theoretical basis for choosing this filtering strategy.

Second, the EKF is used for VIO state prediction, which recursively estimates state vectors, combines IMU high-frequency motion measurements and visual observation constraints to perform observation updates and 3D coordinate calculations, and then realizes the optimal estimation of the UAV's 3D position.

Finally, the UAV state evolution is driven by the IMU measurements, whose discrete-time state transfer model is Equation (1), while the binocular camera provides the observations Equation (2), and the corresponding observations and equations are Equation (3), respectively.

$$\mathbf{X}_{t+1} = f(\mathbf{X}_t, \mathbf{u}_t) + \mathbf{w}_t. \quad (1)$$

$$\mathbf{z}_t = h(\mathbf{X}_t) + \mathbf{v}_t. \quad (2)$$

$$h(\mathbf{X}_t) = \mathbf{p}_t + \mathbf{R}_c \mathbf{t}_c. \quad (3)$$

Ultimately, the 3D position estimation of the UAV is obtained through the prediction and updating of the EKF as in Equation (4). The Kalman gain K_t ensures the optimal fusion of visual and IMU information. The computational complexity $O(n^3)$ of the EKF is determined by the state covariance matrix update and Kalman gain computation, where n is the state dimension.

$$\mathbf{p}_t^+ = \mathbf{p}_t^- + \mathbf{K}_t(\mathbf{z}_t - h(\mathbf{X}_t^-)). \quad (4)$$

The EKF method can realize efficient local fusion with low computational cost and provide accurate 3D position information for subsequent global correction and distributed sensing sharing.

3.1.3. Global Calibration

GPS and UWB are introduced as global correction sensors in the global correction phase. Absolute and relative position information is provided to construct the absolute position constraints under the unified coordinate system. GPS provides global absolute position and UWB provides relative distance information between UAVs.

The GPS and UWB observations are defined in Equation (5), where $\mathbf{z}_t$ is the sensor observation, including GPS position and UWB ranging, $h(\mathbf{X}_t)$ is the observation equation, which transforms the system state to the GPS/UWB observation space, and $\mathbf{v}_t$ is the measurement noise.

$$\mathbf{z}_t = h(\mathbf{X}_t) + \mathbf{v}_t. \quad (5)$$

The GPS observation equation is $\mathbf{z}_t^{\text{GPS}} = \mathbf{p}_t^{\text{GLOBAL}} + \mathbf{v}_t^{\text{GPS}}$. $\mathbf{p}_t^{\text{GLOBAL}}$ is the 3D coordinates of the globally estimated state. The relative ranging information provided by the UWB is shown in Equation (6).

$$\mathbf{z}_t^{\text{UWB}} = \|\mathbf{p}_t^{\text{GLOBAL}} - \mathbf{p}_t^{\text{GLOBAL, ref}}\| + \mathbf{v}_t^{\text{UWB}}. \quad (6)$$

In addition, the global position is used to achieve consistent multi-UAV information through the Labeled Multi-Bernoulli Filter (LMB). GPS and UWB data are also used to ensure the accuracy and robustness of the global fusion results as a complement to the

relative position information between multiple UAVs. The global state is updated as in Equation (7).

$$\mathbf{X}_t^{\text{GLOBAL}} = \mathbf{X}_{t-1}^{\text{GLOBAL}} + \Delta \mathbf{X}_t. \quad (7)$$

This means that the localization $\mathbf{X}_t^{\text{GLOBAL}}$ of the current moment is calculated through the state $\mathbf{X}_{t-1}^{\text{GLOBAL}}$ of the previous moment and the positional transformation $\Delta \mathbf{X}_t$ between the current moment and the previous moment.

3.1.4. LSTM Network Building and Training

The LSTM network model designed in this section learns the temporal variation rules of VIO localization, ultra-wideband ranging, and GPS localization data through multi-input structure and temporal modeling capability. The temporal dependencies in different sensor data streams are mined to achieve efficient fusion of multi-sensor data and globally consistent localization. The LSTM training phase includes the following three main steps:

1. Forward computation of the output value of each neuron, for LSTM, i.e., $\mathbf{f}_k$, $\mathbf{i}_k$, $\mathbf{C}_k$, $\mathbf{o}_k$, $\mathbf{h}_k$. The values of the five vectors. The input temporal data $\mathbf{x}_k = [\mathbf{x}_{k,m+1}, \mathbf{x}_{k,m+2}, \dots, \mathbf{x}_{k,l}]$ into the LSTM network, the timing information is processed through the hidden layer of the network and the memory unit, and the estimated value of the bit position at the current moment is output. The bit position estimation value at the current moment can be obtained by the following equation:

- a. Forget Gate: Determines which historical information to discard.

$$\mathbf{f}_k = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{k-1}, \mathbf{x}_k] + \mathbf{b}_f). \quad (8)$$

- b. Input Gate: Determines which new information to update.

$$\mathbf{i}_k = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{k-1}, \mathbf{x}_k] + \mathbf{b}_i). \quad (9)$$

- c. Candidate Memory Cell: Generates a new candidate memory.

$$\tilde{\mathbf{C}}_k = \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{k-1}, \mathbf{x}_k] + \mathbf{b}_C). \quad (10)$$

- d. Memory Cell Update: Updates the state of the memory cell.

$$\mathbf{C}_k = \mathbf{f}_k \odot \mathbf{C}_{k-1} + \mathbf{i}_k \odot \tilde{\mathbf{C}}_k. \quad (11)$$

- e. Output Gate: Determines which information to output.

$$\mathbf{o}_k = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{k-1}, \mathbf{x}_k] + \mathbf{b}_o). \quad (12)$$

- f. Hidden State Update.

$$\mathbf{h}_k = \mathbf{o}_k \odot \tanh(\mathbf{C}_k). \quad (13)$$

2. Backpropagate the error term δ for each neuron. The backpropagation of LSTM consists of two directions: one is to backpropagate the error at each moment along the time, and the other is to propagate the error to the next level. Since the output is computed by the activation function $f(\cdot)$, the error term δ needs to consider the derivative of the activation function.

$$\sigma(z) = y = \frac{1}{1 + e^{-z}}, \quad (14)$$

$$\sigma'(z) = y(1 - y), \quad (15)$$

$$\tanh(z) = y = \frac{e^z - e^{-z}}{e^z + e^{-z}}, \quad (16)$$

$$\tanh'(z) = 1 - y^2. \quad (17)$$

Let the activation function of the gate be the sigmoid function (14), the activation function of the output be the tanh function (15), and the derivatives of the sigmoid and tanh functions be (16) and (17).

3. The gradient of each weight is computed based on the corresponding error term. The backpropagation utilizes the chain rule combined with the activation function derivatives of each neuron to calculate the error term delta and weight update layer by layer starting from the output layer. To ensure the training effect of LSTM network, the loss function is designed to balance the local variation and global constraints. Mean Square Error (MSE) is chosen as the loss function, and the mathematical representation of MSE is shown in Equation (18).

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (18)$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and N is the sample size.

$$L_{local} = \frac{1}{N} \sum_{i=1}^N \|\Delta T_{i \rightarrow j}^{pred} - \Delta T_{i \rightarrow j}^{VIO}\|^2, \quad (19)$$

$$L_{global} = \frac{1}{M} \sum_{i=1}^M \|P_t^{pred} - P_t^{GPS}\|^2, \quad (20)$$

$$L = \alpha L_{local} + \beta L_{global}. \quad (21)$$

The loss function is categorized into local change constraint loss L_{local} (19) and global position constraint loss L_{global} (20). Where $\Delta T_{i \rightarrow j}^{pred}$ is the local positional change output by LSTM. $\Delta T_{i \rightarrow j}^{VIO}$ is the observed position change provided by VIO. P_t^{pred} represents the global position output by LSTM. P_t^{GPS} is the observed position provided by GPS. The total loss function is Equation (21).

where α and β are loss weights that control the effects of local versus global constraints. Learnable W_{local} and W_{global} are introduced in the fusion layer of LSTM for dynamically adjusting the importance of local and global information. The local and global fusion is shown in Equation (22).

$$h_t = W_{local} \cdot h_t^{local} + W_{global} \cdot h_t^{global}. \quad (22)$$

In the output phase after training, the final prediction of the current moment's global positional pose $T_t^{pred} = [x, y, z, q_w, q_x, q_y, q_z]$ is output at the LSTM output layer. And the expression for the output global positional pose prediction is Equation (23).

$$T_t^{pred} = f_{out}(h_t). \quad (23)$$

The output layer converts the fused state information into the final localization result of the UAV. The 3-degree-of-freedom position information of the UAV, i.e., position (x, y, z) , is output. The output results are compared with the global constraints through a feedback mechanism to correct the localization error and improve the

system robustness. Local computation focuses on high dynamic accuracy in a short period of time, while global computation ensures long-term positioning consistency, and the combination of the two can realize the best balance between accuracy and real-time performance.

3.2. LMB-Based Distributed Cooperative Localization Algorithm for Multiple UAVs

3.2.1. Labeled Multiple Bernoulli Algorithm

The Labeled Multi-Bernoulli Filter (LMB Filter) algorithm is one of the important advances in the field of multi-target tracking. First proposed by Reuter et al. in 2014 [30], the algorithm is able to provide a unique identity label for each target while striking a balance between computational complexity and accuracy, which effectively extends and improves the traditional Multi-Bernoulli Filter (MB Filter). The traditional RFS-based filter avoids the complex data correlation problem in multi-target tracking by uniformly processing the set of targets, but fails to retain the identity information of the targets. B-T Vo's team [31] in proposed to introduce the labeling space into the stochastic finite set to derive the potential-balanced multi-bernoulli filter with unbiased estimation.

Label Random Finite Set (Label RFS) introduces a discretized label space $\mathbb{L} = \{a_i : i \in \mathbb{N}\}$, with the range of values of the label variable a_i denoted as the set of positive integers N . The multi-Bernoulli distribution, which originally described only the single-targeted state x , is extended to contain the label's information $(x, l) \in \mathbb{X} \times \mathbb{L}$. As a result, in the space $\mathbb{X} \times \mathbb{L}$ constituting the Label RFS, one can establish a mapping relation $\mathbb{X} \times \mathbb{L} \rightarrow \mathbb{L}$, denoted as $\mathbb{L}((x, \ell)) = \ell$.

The LMB filtering algorithm has a great advantage in dealing with target state estimation and identification at the same time. It has been widely used in the fields of UAV cluster cooperative target tracking, intelligent traffic monitoring, and military target surveillance. The probability density function of the Labeled Multiple Bernoulli (LMB) distribution can be expressed as Equation (24).

$$\pi(X) = \Delta(X)w(\mathbb{L}(X))p^X, \quad (24)$$

where $\Delta(X)$ is the label indicator.

$\Delta(X)$ takes the value 1 if and only if the target state corresponds one-to-one with the label, otherwise $\Delta(X) = 0$. Ensure that each target state corresponds accurately with its corresponding label. The other parameters in Equation (24) are expressed as Equations (25) and (26).

$$w(L) = \prod_{i \in \Sigma} (i - r^{(i)}) \prod_{\ell \in L} \frac{1_{\mathcal{L}}(\ell)r^{\ell}}{1 - r^{(\ell)}}, \quad (25)$$

$$p(x, \ell) = p^{\ell}(x). \quad (26)$$

3.2.2. Multiple Bernoulli Fusion Algorithm for Locally Consistent Labels

The sensing capability of individual UAVs is limited, and UAV clusters with distributed networking are often constructed to fuse the local sensing results of each UAV to derive the global sensing information. However, under the limitation of communication network, each node can only obtain the information of the local neighboring UAVs, and cannot access the data of the whole cluster. This not only makes the local observation different, but also prone to information inconsistency and synergistic dissonance when fusion occurs. To realize global sensing and global localization of distributed networking UAV clusters, two key issues need to be solved.

First, there is no unified central node in the cluster, and UAVs can only exchange data with their nearest neighbor communication nodes, making it difficult to directly

aggregate global data. It can only fuse the information with the local sensing results of neighboring nodes to ensure that the local data reflects a consistent global posture after multiple iterations of fusion.

Second, the data correlation of different UAVs is complex when restricted communication is used, and the public information cannot be accurately assessed. Traditional data fusion methods may result in distorted results due to redundancy of information, affecting system synergy and decision-making quality.

A distributed data fusion algorithm named Local Consistency Labeled Multi-Bernoulli Fusion Algorithm is proposed to address the above core problem. The algorithm is constructed based on the consistency iteration mechanism, network weight fusion criterion and Labeled Random Finite Set (LRFS) theory. In the specific application, the algorithm first allows each UAV to independently compute the local sensing results in the form of LMB-RFS. The following steps are then performed sequentially to achieve the goal of information consistency and cooperative fusion under communication constraints.

1. Local sensing.

In a UAV cluster with distributed networking, each UAV i has a unique label and computes a local sensing result in the form of LMB-RFS using its own sensors. The result not only includes the probabilistic description of the target state, but also records and stores its corresponding label information as $\pi^i = (r^i, p^i)$, which serves as the basic data for subsequent information exchange.

2. Information exchange and initial consistency $\bar{r}_1$ computation.

Within a given number of consistency iterations C , each UAV performs information transfer with nodes in its neighborhood. During the $c \in C$ iteration, two sub-steps are performed.

- a. Information exchange.

Any UAV i in the cluster sends the localization $r_{1,c-1}^i$ obtained after the last iteration to its neighbors, with $r_{1,0}^i = r_1^i$.

- b. Consistency calculation.

After receiving the data from the neighboring machines, each node weights the local results based on the predefined consistency weights to correct the local results, so that the local results gradually converge to the global average state. Denote Ω as the consistency matrix whose i th row and j th column elements are ω_{ij} . Based on the information received from other UAVs, the labeled multi-Bernoulli algorithm ensures the consistency of the states of all UAVs by correcting the previous state estimates.

If there is a large error in the estimation of the state of a particular UAV, the other UAVs will correct the error by fusing the information, ultimately keeping the state of all UAVs in the system consistent. Calculate $r_{1,c}^i = \sum_{j \in N^i} \omega_{ij} r_{j,1,c-1}^j$, $\forall i \in N$. where ω_{ij} denotes the consistency weight, satisfying $\omega_{ij} > 0, \forall i, j \in N$ and $\sum_{j \in N^i} \omega_{ij} = 1, \forall i \in N$.

3. Secondary information exchange with consistent iterative updating $\bar{q}_1$.

The labeled multi-Bernoulli algorithm makes corrections based on previous state predictions after receiving data from other UAVs. For each update, the algorithm calculates the probability of all targets and weights the predicted position of each UAV based on the calculated consistency weights. If the UAV's state estimate does not match that of its neighbors, the system will use relative positioning to make corrections, thus reducing the propagation of errors and inconsistencies. Given the number of coherence iterations C . During the $c \in C$ iteration, two sub-steps are performed.

- a. Information exchange.
Any UAV i in the cluster passes its result $q_{1,c-1}^i$, after the previous iteration, to other UAVs in the neighborhood with $q_{1,0}^i = q_1^i$.
- b. Consistency iteration.
According to the consistency weight defined by ω_{ij} in Equation (27), the weight satisfies the non-negativity $\omega_{ij} > 0$ for all nodes $\forall i, j \in N$ and satisfies the normalization condition in Equation (28). After iteration, the fusion result of each node will converge to the overall mean, reaching the calculation of the global fusion parameter $\bar{q}_1$ described in Equation (29).

$$q_{l,c}^i = \sum_{j \in N^i} \omega_{ij} q_{j,l,c-1} \quad \forall i \in N, \quad (27)$$

$$\sum_{j \in N^i} \omega_{ij} = 1 \quad \forall i \in N, \quad (28)$$

$$q_{l,c}^i = \sum_{j \in N^i} \omega_{ij} q_{j,l,c-1} \quad \forall i \in N. \quad (29)$$

4. Global perceptual fusion output.

The whole process is a dynamic iterative process. Over time, each UAV continuously utilizes the local sensor data to update its localization with information from other UAVs. Meanwhile, the LSTM network continuously optimizes the sensor data fusion process, and the labeled multi-Bernoulli algorithm globally ensures that the positioning of each UAV converges. Each UAV is iteratively updated based on the information from its neighbors, gradually converging to a consistent state. This approach not only reduces the large-scale inconsistency caused by individual UAV errors, but also gradually corrects the local errors through multiple rounds of iterations. From the consistency principle, it can be seen that under the distributed networking conditions, after many rounds of iteration, the global sensing fusion results will definitely realize the collective consensus. The complete flowchart of the proposed co-localization algorithm is summarized in Figure 2.

3.3. Design of the SE-LSTM-SA Localization Module

LSTM-Attention network is a deep learning model that combines long short-term memory (LSTM) network and attention mechanism. In recent years, LSTM-Attention network has significant advantages in the fields of time series prediction and multimodal data classification (see Figure 3 for a typical structure.) The LSTM network relies on the gating mechanism to solve the problem of gradient disappearance in long series data, and the attention mechanism can dynamically assign weights and highlight key information to improve the performance [32].

Squeeze-and-Excitation (SE) is an attention mechanism for channel-level features proposed by Hu et al. in 2018 [33]. The core idea is to increase the weight of important features by explicitly modeling the dependencies between channels. At the same time, unimportant features are suppressed and the channel feature responses are adaptively recalibrated. This enhances the model's ability to extract key features, and since its proposal, it has been studied and practiced in the fields of performance prediction [34] and recognition classification [35]. Figure 4 illustrates the framework of the SE module.

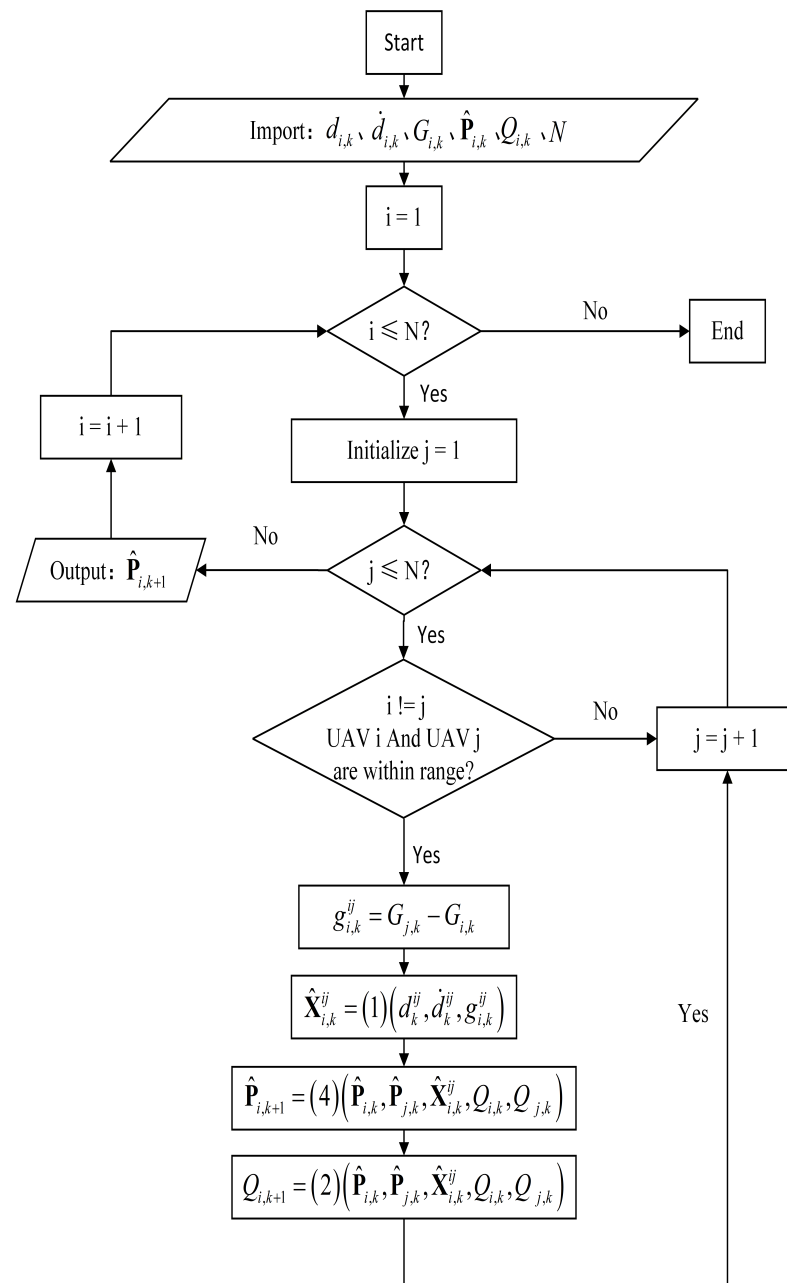


Figure 2. Co-localization algorithm process.

The compression incentive mechanism mainly consists of two stages, Squeeze and Excitation. In the Squeeze phase, the spatial dimension of each channel is compressed into a scalar by the Global Average Pooling (GAP) method. The compressed global spatial information generates channel descriptors, which can capture the global context information and reduce the interference of spatial dimension on feature extraction. Formally embodied as a reduction U in the spatial dimension $H \times W$, the 3D spatial mapping features of size $C \times H \times W$ for each channel are compressed into a real number of $C \times 1 \times 1$, where the c th element of the statistic z is computed by Equation (30).

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j). \quad (30)$$

In the Excitation operation, the dependencies between different channels are learned using the fully connected layer and the nonlinear activation function to assign channel

weights. The channel features are then recalibrated by weighted summation. The sigmoid activation function is used to map the output to the range $[0, 1]$. Enhance important features and suppress redundant features. The specific calculations are shown in Equation (31).

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \quad (31)$$

where σ denotes the Sigmoid activation function and δ denotes the modified linear unit ReLU.

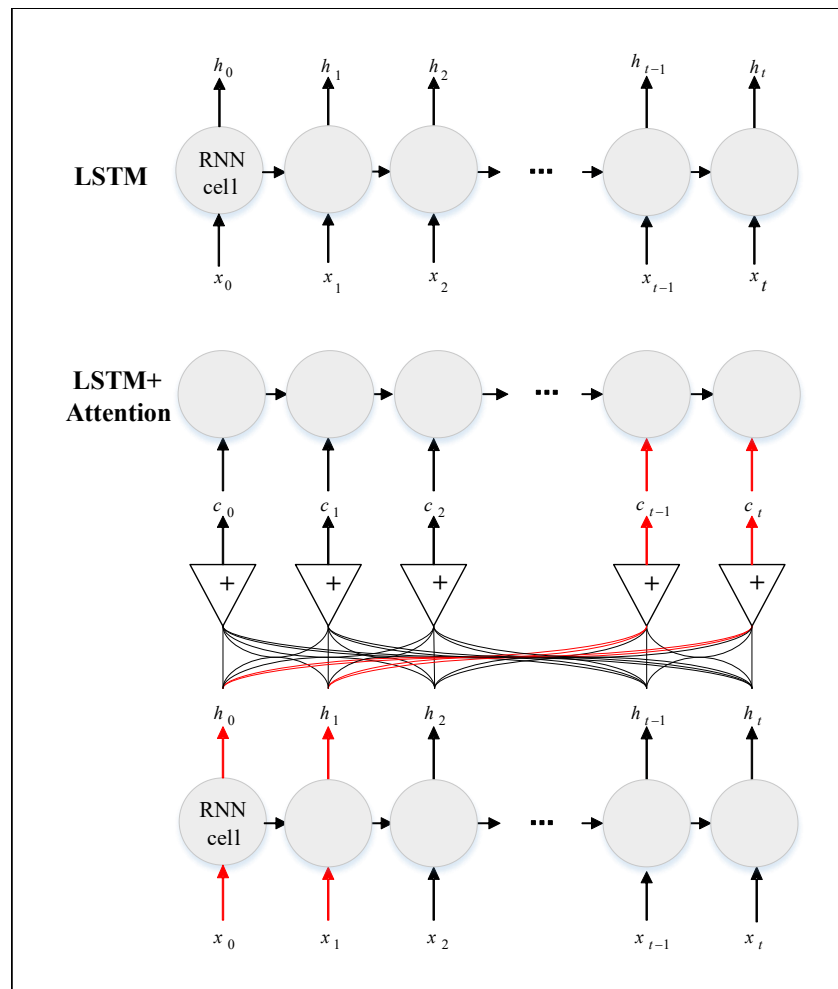


Figure 3. LSTM and LSTM systems with attention mechanisms.

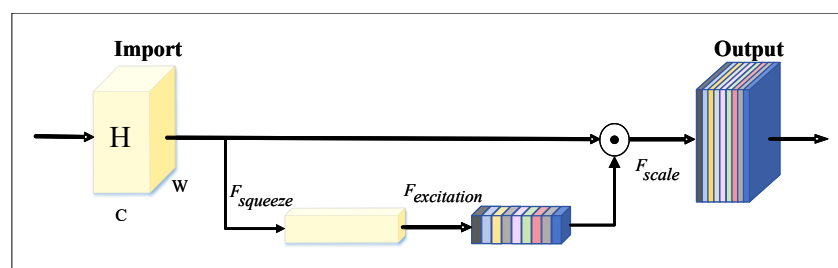


Figure 4. SE module. The different colors in the output blocks and the excitation vector represent different feature channels and their correspondingly recalibrated weights.

Finally, there is the operation of Reweight. By multiplying the channel-by-channel weighting and mixing the channel weights with the original features, the recalibration of the original features in the channel dimension is completed. See Equation (32) for details.

$$F_{scale}(u_c, s_c) = s_c \cdot u_c. \quad (32)$$

Introducing a compression excitation layer in the LSTM network can effectively improve the feature extraction ability of the model for multi-sensor data. Unlike temporal self-attention, which focuses on inter-time-step dependency modeling, the SE mechanism explicitly performs channel-wise feature recalibration. This is particularly important in heterogeneous multi-sensor fusion because different sensing modalities exhibit varying reliability under dynamic interference conditions. The compressed excitation layer enhances the model's sensitivity to key information by selectively enhancing highly informative sensing channels while suppressing degraded ones, thus significantly improving the localization accuracy and robustness. The compressed excitation module is combined with LSTM to form the compressed excitation LSTM network model shown in Figure 5.

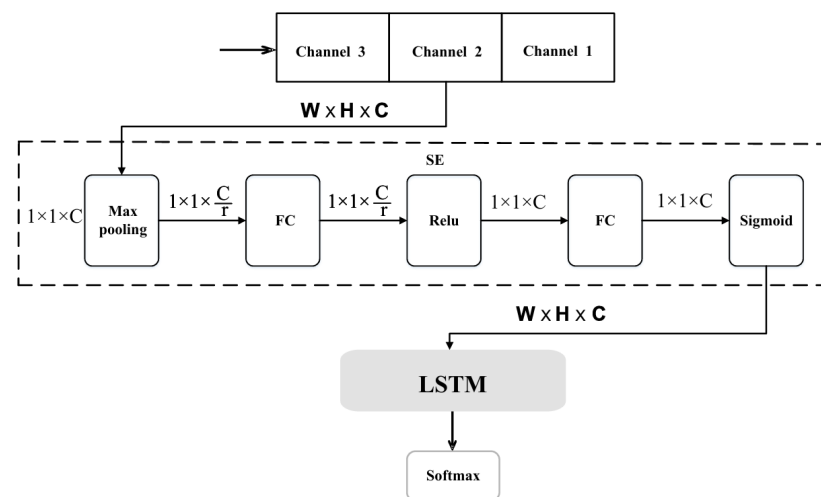


Figure 5. SE-LSTM network.

In the multi-UAV temporal trajectory localization task, the module order design of the network structure has an important impact on the model performance. A reasonable module order can ensure that the model can effectively extract features and capture timing dependencies, so as to realize accurate trajectory prediction. The adopted network model structure is shown in Figure 5. The model structure design systematically considers feature optimization, timing modeling, information filtering and output mapping. It ensures that the model can efficiently handle complex multi-UAV trajectory data.

In terms of data loading and model training, the network is trained offline using pre-processed and fused multi-sensor data containing 99,951 samples, divided into 107 training batches and 19 test batches. The input data shape is [batch_size, seq_len, input_channels], where batch_size is the number of samples in each batch, seq_len is the time step length of each sample, and input_channels is the feature dimension of each time step. The network parameters are optimized using the Adam optimizer with a dynamic learning rate decay strategy to accelerate convergence and avoid local optima. The initial learning rate is empirically set to 0.001, and the loss weights α , β , and λ are fine-tuned iteratively based on the training performance.

The model weights and regulates the input data channels through the adaptive excitation module. The input features are 21-dimensional data, which are compressed by

adaptive average pooling to compress the time-step dimension, and then the channel weights are adjusted by linear mapping and activation function. Dependencies between features are extracted to enhance useful information and suppress redundant information.

After the adaptive excitation module, the data enters the bidirectional long and short-term memory network layer. The input dimension is 21 and the number of unidirectional hidden states is 256, forming a 512-dimensional feature representation. Ensuring the full extraction of time-dependent information and taking into account the pre- and post-textual information helps to capture the complex dynamic patterns in the UAV motion trajectory. dropout strategy is introduced into the LSTM to alleviate the overfitting problem and enhance the generalization ability.

The high-dimensional features based on the LSTM output enter the self-attention mechanism layer. Queries, keys and values are generated by three independent linear transformations, all mapping 512-dimensional features to the same dimension. The Softmax function is utilized to obtain the attention weights at each moment, and the relationships between different time steps are weighted and summed. The importance of global information is highlighted to compensate for the information loss problem of traditional LSTM in capturing remote dependencies.

Finally, it is processed by dropout regularization. The model utilizes a fully connected layer to map the 512-dimensional features extracted by the self-attention mechanism to a 9-dimensional output, corresponding to the position estimation of the three UAVs in 3D space. A loss function incorporating consistency constraints is designed during the training process as in Equation (33). The multi-UAV prediction results are smoothed using the weight function and the LCWU weighted position update method to realize the inter-node consistency characteristic constraints.

$$\text{Total Loss} = \text{MSE Loss} + \lambda \times \text{Consistency Loss}. \quad (33)$$

The designed network structure optimizes the input features through the SE module, the LSTM layer captures the timing dependencies, the attention layer focuses on the critical time steps, and the fully connected layer implements the feature-to-task output mapping. This enhances the localization accuracy of multi-UAV timing trajectories, improves robustness and generalization, and provides an effective solution for cooperative localization in complex dynamic environments.

3.4. Consistent Localization Implementation Based on Compression-Excited LSTM-SA Networks

The SE-LSTM-SA network and the LCWU-LMB distributed fusion module operate in a hierarchical collaborative manner. Specifically, the SE-LSTM-SA network performs local temporal sensor fusion and outputs refined single-UAV state estimates, including position and motion features under heterogeneous sensor observations. These locally optimized state estimates are then treated as probabilistic observations by the distributed LCWU-LMB framework, which further performs inter-UAV consistency fusion through iterative neighbor-based information exchange. Therefore, the proposed framework establishes a two-level cooperative localization architecture: (1) local adaptive heterogeneous sensor fusion through SE-LSTM-SA; and (2) global distributed consistency maintenance through LCWU-LMB.

Building upon this hierarchical architecture, the proposed multi-UAV coherent localization algorithm incorporates the Local Consistency Weight Update (LCWU) mechanism based on Labeled Random Finite Set (LRFS) theory. The overall structure of the model consists of an adaptive excitation module, a bi-directional long short-term memory network, a self-attention mechanism, and a fully connected mapping layer. It is able to capture the long-range dependency while ensuring the integrity of the timing information, and

correct the consistency of the prediction results of each UAV. The algorithm flow is shown in Algorithm 1.

The data fusion model constructed through the LSTM network is described as Equation (34).

$$f_{SE-LSTM-Attention}(\hat{X}_{k-m}, \dots, \hat{X}_{k-1}, \hat{X}_k, \Theta_{SE-LSTM-Attention}). \quad (34)$$

where $\Theta_{\text{Attention-LSTM}}$ denotes the set of all trainable weights and biases.

$\Theta_{\text{Attention-LSTM}}$ is specified as Equation (35), where l denotes the number of layers of the LSTM network.

$$\Theta_{SE-LSTM-Attention} = \left\{ \left\{ W_f^l, W_i^l, W_o^l, W_c^l, b_f^l, b_i^l, b_o^l, b_c^l \right\}_{l=1}^N, W_{out}, b_{out} \right\}. \quad (35)$$

Algorithm 1 SE-LSTM-Attention Distributed Cooperative Localization Algorithm

Require: VIO, GPS sequence $\{P_t, P_{t+1}, \dots, P_{t+M-1}\}$, UWB sequence $\{d_t, d_{t+1}, \dots, d_{t+M-1}\}$

1: Number of training epochs n , initial consistency matrix $\Omega = [\omega_{ij}]_{n \times n}$

Ensure: Weight parameters $\Theta_{SE-LSTM-Attention} = \{W_f^l, W_i^l, W_o^l, W_c^l, b_f^l, b_i^l, b_o^l, b_c^l\}_{l=1}^N, W_{att}, b_{att}$

2: Data fusion model $f_{SE-LSTM-Attention}(\hat{X}_{k-m}, \dots, \hat{X}_{k-1}, \hat{X}_k, \Theta_{SE-LSTM-Attention})$

3: Trajectory prediction sequence of the last epoch, calculated errors RMSE, MaxPE, MAE

4: Initialize weights and biases

5: SE layer weight optimization for heterogeneous channel calibration

6: Forward propagation: Output sequence values through the designed LSTM network layer and attention layer

7: Calculate total loss function = MSE loss + $\lambda \times$ consistency loss

8: Backpropagation and parameter update: Calculate the gradient of each node based on Adam algorithm, update parameters

9: LCWU weight update **▷ Cooperative weight adjustment via neighbor interaction**

10: Update consistency matrix **▷ Distributed consensus state exchange**

11: Consistency Labeled Multi-Bernoulli fusion **▷ Cooperative multi-UAV fusion**

12: Repeat forward propagation, loss function calculation, backpropagation, parameter update, LCWU weight update, consistency matrix update and consistency label multi-Bernoulli fusion until the loss value converges or reaches the preset number of iterations

13: Save model parameters

14: **Prediction Phase (Online Deployment)**

Require: Time series at M moment

Ensure: Position estimate at K moment

15: Use the saved model parameters to predict the position estimate at k moment, and output the calculated errors RMSE, MaxPE, MAE

Hardware Deployment, Inference Time, and System Scalability

To address the real-time execution requirements in multi-UAV systems, it is essential to distinguish between the training and deployment phases. The SE-LSTM-SA network is trained offline on a standard computing workstation. Once trained, the lightweight model parameters are deployed online for real-time inference on UAV-mounted embedded edge computers, such as the NVIDIA Jetson series (e.g., Jetson Nano or Xavier NX), which provide sufficient tensor processing capabilities for deep learning inference.

Based on our evaluation using the validation dataset (1669 data frames), under offline-trained online inference settings on an NVIDIA Jetson Xavier NX platform, the average forward inference latency of the neural localization module is approximately 0.63 milliseconds per data frame, excluding inter-UAV communication delay. This extremely low computational latency proves that the algorithm can easily satisfy the rigorous real-time self-localization constraints of highly dynamic aerial missions.

The computational complexity of the proposed framework is dominated by the LSTM temporal modeling and self-attention operations. Since only local-neighbor information exchange is required in the LCWU mechanism, the communication overhead grows linearly with the number of neighboring UAVs rather than the total swarm size, ensuring scalability under communication-constrained environments. Furthermore, regarding the scalability of the proposed distributed cooperative framework, the minimum number of UAVs required to execute this algorithm is two. The maximum number of UAVs in the cluster is practically limited by the local communication bandwidth and the wireless ad-hoc networking protocol's capacity. Typically, the framework supports a local sub-swarm of 10 to 15 UAVs cooperating seamlessly within a shared communication range.

4. Simulation Experiments and Results

4.1. Experimental Setup

Vision sensors, IMU, GPS, and UWB ranging sensors are selected and fused for UAV-related tasks. Single sensor limitations can be overcome to achieve highly accurate and stable positioning in complex environments and mission requirements.

Due to the differences in sampling rates, acquisition moments, and installation locations of different types of sensors, the direct use of unprepared data can easily lead to the accumulation of errors, so it is necessary to perform time alignment before data fusion. Commonly used time alignment methods include parameter estimation such as least squares, Kalman filtering, and interpolation such as interpolation and extrapolation, curve interpolation, curve fitting, and so on. In the XTDrone simulation platform [36], the global start or reference time can be set, and the `numpy.interp` function can be used to linearly interpolate the VIO data to realize the time alignment.

The position and attitude of the UAV in 3D space are collectively referred to as the position information, which is described by a six-dimensional vector containing 3D position (x, y, z) and 3D rotation (ψ, ϕ, θ) . The experiment needs to transform the measurement information of each sensor into the same spatial coordinate system through the coordinate transformation matrix. According to the installation and measurement mode of the sensors, the corresponding conversion matrix is designed, and the coordinate system of the machine body is used as the main coordinate system to realize the unification of the coordinate system.

The validation experiments were conducted on the ROS-based XTDrone simulation platform running on Ubuntu 20.04. To verify the feasibility of the algorithm on restricted embedded devices, the localization network was trained offline using PyTorch 2.2 and deployed online on an NVIDIA Jetson Xavier NX computing platform. To emulate realistic multi-sensor configurations, the sensor data acquisition rates were strictly defined: the IMU frequency was set to 100 Hz, the camera operated at 30 Hz, GPS measurements were updated at 5 Hz, and UWB ranging data were obtained at 10 Hz. The complete dataset comprises 99,951 samples, which were partitioned into 107 training batches and 19 testing batches to thoroughly evaluate the network's generalization capability.

4.2. Results and Discussion

Based on the simulation experimental data, the model is analyzed to explore the model performance and effectiveness. Figure 6 presents the trend of training loss and testing loss of the model over 50 training epochs. It can be seen that the training and testing losses decrease significantly as the training progresses.

As shown in the figure, the initial training loss of 0.10197 is reduced to 0.00024 after 50 epochs of training, which reflects the good convergence of the model. In the first stage, the training loss drops to 0.00157 within the first 10 epochs, and the testing loss drops to

0.00251. This is the key stage for capturing the dynamic laws and rapidly reducing errors. The model optimizes stably around the 30th epoch, and the loss decreases gradually. In the later stage, the 50th epoch reaches complete convergence, and the test loss decreases from 0.00861 to 0.00021 and stabilizes. In the experiment, the difference between the training and testing losses tends to stabilize, and their proximity indicates strong generalization ability without overfitting.

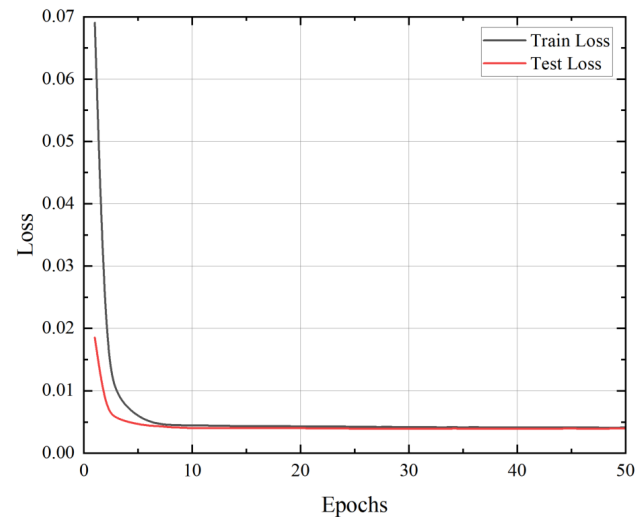


Figure 6. Loss Variation During the Iterative Process.

To comprehensively evaluate the performance of the proposed learning-based fusion method, Figure 7 illustrates the 3D trajectory comparison among the Ground Truth, the traditional Extended Kalman Filter (EKF) predicted trajectory, and the LSTM-predicted trajectory.

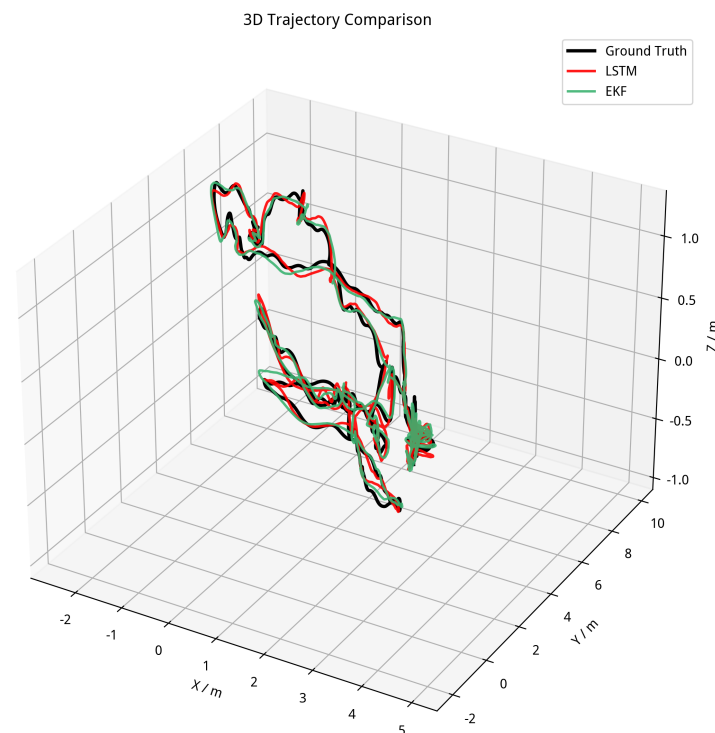


Figure 7. 3D Trajectory Comparison among Ground Truth, EKF, and LSTM frameworks.

The experimental results demonstrate that the LSTM network trajectory exhibits a significantly higher overlap with the Ground Truth, indicating improved adaptability in complex dynamic environments. In contrast, the EKF-filtered trajectory shows noticeable deviations in several regions. Specifically, its prediction accuracy degrades significantly during sudden changes in the target's motion state. This primarily occurs because traditional EKF methods rely heavily on predefined linear approximations, making them highly vulnerable to rapid error accumulation under unpredictable, highly dynamic maneuvers. The data-driven LSTM approach successfully overcomes this by accurately modeling long-term non-linear temporal dependencies.

To further quantify the localization accuracy, Table 1 presents a comprehensive quantitative comparison of the prediction errors among the traditional EKF, the baseline LSTM, and the proposed self-attention-enhanced localization models. As shown in the table, the learning-based algorithms consistently outperform the traditional EKF approach across all evaluation metrics, providing strong quantitative evidence for their superior stability and accuracy in dynamic environments.

Table 1. Quantitative comparison of localization errors between EKF and LSTM.

Method	RMSE (m)	MSE	MAE (m)	MaxPE (m)
EKF	0.0984	0.0097	0.0867	0.2614
LSTM	0.0269	0.0048	0.0325	0.1199

Furthermore, to verify the directional stability of the model, the Root Mean Square Error (RMSE) across the X, Y, and Z axes during the training and testing phases was recorded, as shown in Figure 8. The localization RMSE in all directions rapidly converges and stabilizes at approximately 0.04 m during the testing phase, indicating high prediction accuracy and strong generalization capability across all spatial dimensions.

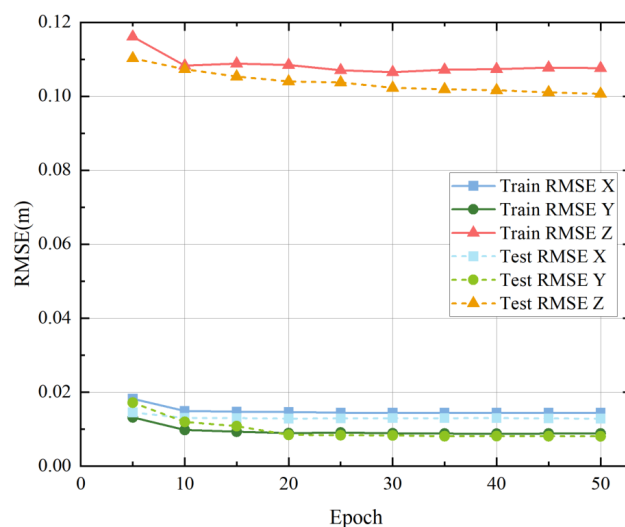


Figure 8. Training and Testing RMSE across X, Y, and Z directions.

Following the baseline validation, ablation experiments are conducted to evaluate the localization errors of different attention-enhanced models. Figure 9 compares the training loss and test loss variations of the three models: traditional LSTM, LSTM-SA, and Squeeze-excited LSTM-SA (SE-LSTM-SA), over 50 iterations. All models are trained on the same dataset using identical initial learning rates and decay strategies.

By comparing the training and test losses, it can be seen that after the introduction of the self-attention mechanism, the loss curves present a smoother trend, effectively reducing

the overfitting phenomenon. Compared with LSTM and LSTM-SA, the SE-LSTM-SA shows particularly significant improvement in the test loss values, with faster convergence and smoother curves. This confirms that the Squeeze-and-Excitation mechanism enhances the representational ability of the model by selectively recalibrating heterogeneous sensor channels. This improvement indicates that the SE mechanism effectively suppresses unreliable sensor features and adaptively enhances informative channels under dynamic noise conditions, thereby improving feature representation quality before temporal fusion.

The quantitative evaluation results of the ablation experiments are illustrated in Figure 10, alongside the quantitative data detailed previously in Table 1. As a baseline model, the traditional LSTM shows significant errors in predicting all assessment metrics of the task. It has a larger error in certain difficult position predictions, indicating that the traditional LSTM model is weak in predicting sudden complex maneuvers.

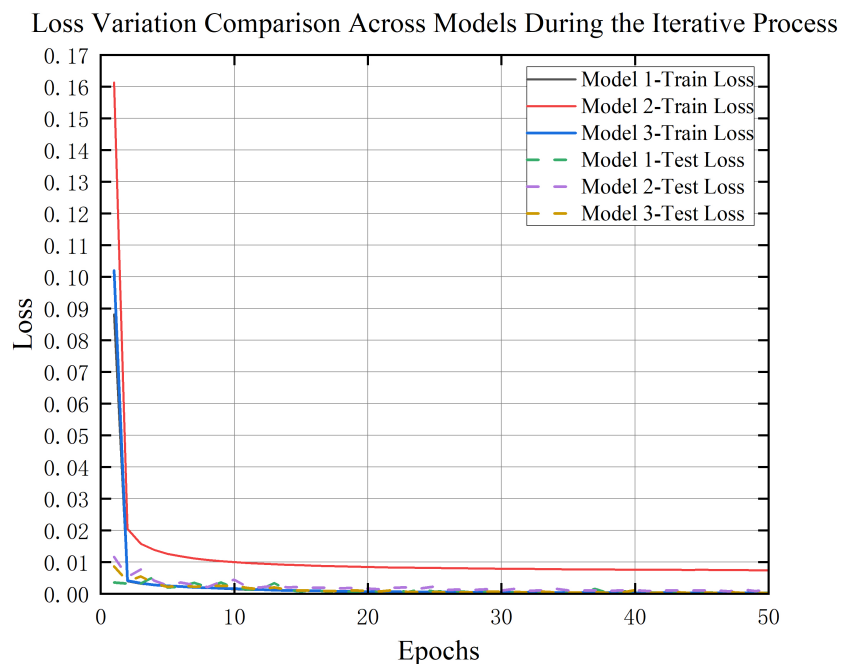


Figure 9. Loss Variation Comparison Across Models During the Iterative Process.

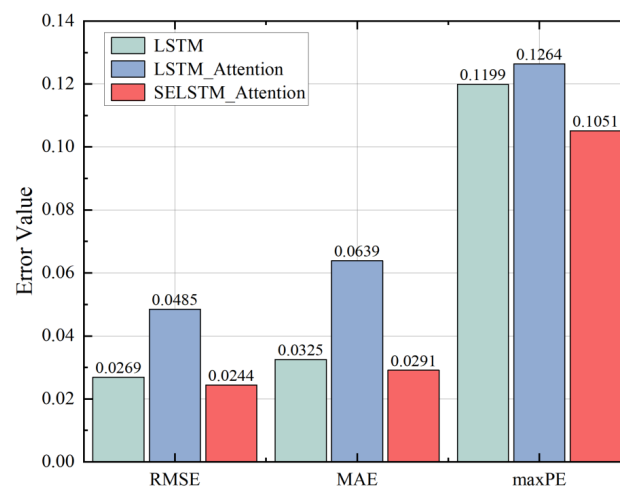


Figure 10. Comparison of error values for different network structures.

$$\text{Relative Improvement} = \frac{\text{Metric}_{\text{Base}} - \text{Metric}_{\text{New}}}{\text{Metric}_{\text{Base}}} \times 100\%. \quad (36)$$

To formally evaluate the inter-UAV coordination, the global consistency error is defined as the average inter-UAV localization deviation across the swarm:

$$E_{cons} = \frac{1}{N} \sum_{i=1}^N \|x_i - \bar{x}\| \quad (37)$$

where x_i denotes the estimated state of UAV i , and $\bar{x}$ is the swarm-wise average estimated state.

Figure 11 illustrates the comparison of the consistency error of different network structures, including global consistency error and single-step consistency error on the training and test sets. The global consistency error of the LSTM model is 0.095, while the global consistency error of the LSTM-SA and the SE-LSTM-SA drop to 0.016 and 0.0134, respectively. Compared with the baseline LSTM model, the proposed SE-LSTM-SA significantly improves localization accuracy, reducing the global consistency error by approximately 85.9%. The reduction in global consistency error further demonstrates that the proposed LCWU consistency optimization effectively constrains inter-UAV state divergence during distributed localization. This change suggests that the performance in maintaining global consistency improves significantly with the self-attention and squeeze-excitation mechanisms. This improvement is particularly important for distributed multi-UAV systems, where localization inconsistency may accumulate and propagate through cooperative state exchange.

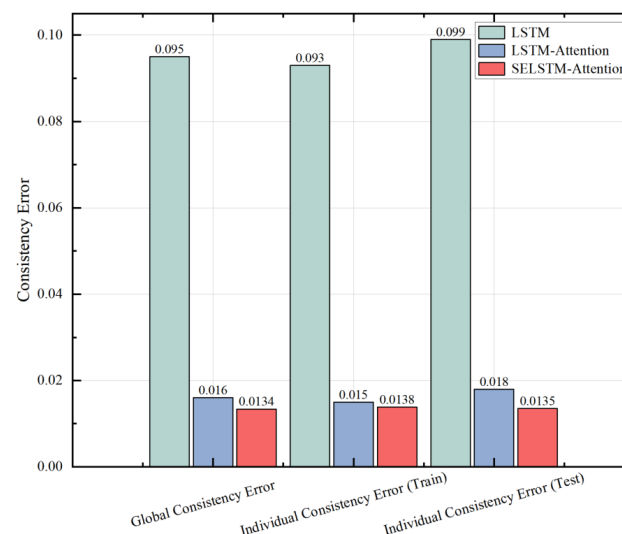


Figure 11. Comparison of consistency error for different network architectures.

In the time dimension, with the introduction of the self-attention mechanism, the global consistency error of LSTM-SA shows faster convergence behavior, indicating that the self-attention mechanism accelerates the convergence process. The SE-LSTM-SA exhibits the smallest convergence factor, indicating that the model can effectively maintain consistency at each time step. The global consistency and single-step consistency convergence times of the SE-LSTM-SA model are noticeably reduced, suggesting that the model converges to the optimal consensus more quickly. This is crucial for reducing computation time and improving response efficiency in practical distributed swarm applications.

By analyzing the single-step consistency error on the training and test sets, the SE-LSTM-SA on the training set is optimal. The remarkably low error on the test set also indicates that the model generalizes exceptionally well to unseen dynamic environments, further validating the performance enhancement brought by the heterogeneous channel-calibrated SE mechanism.

5. Conclusions

This study addresses the distributed cooperative localization problem for multi-UAV systems operating in complex dynamic environments. By integrating heterogeneous multi-sensor fusion, deep temporal feature learning, and consistency-aware distributed state estimation, a robust cooperative localization framework is developed. First, an LSTM-based staged fusion strategy is employed to integrate VIO, GPS, and UWB measurements, thereby improving single-UAV localization accuracy. Subsequently, a Squeeze-and-Excitation LSTM Self-Attention (SE-LSTM-SA) network is introduced to adaptively recalibrate heterogeneous sensor channels and enhance temporal feature extraction under dynamic sensing conditions. Furthermore, a consistency-aware distributed fusion mechanism based on the Labeled Multi-Bernoulli (LMB) framework is incorporated to improve inter-UAV state consistency through local-neighbor information exchange.

Experimental validation conducted on the XTDrone simulation platform demonstrates the effectiveness of the proposed framework. Compared with the traditional Extended Kalman Filter (EKF), the proposed approach significantly improves localization performance, reducing the Root Mean Square Error (RMSE) from 0.0984 m to 0.0269 m while decreasing the global consistency error by approximately 85.9%. In addition, the framework achieves an average inference latency of 0.63 ms per frame on embedded edge computing hardware, demonstrating its feasibility for real-time deployment in distributed multi-UAV systems.

Although the proposed framework achieves promising localization accuracy and consistency performance, several limitations remain. First, the current simulation environment does not fully capture the extreme flight dynamics commonly encountered in aerial gaming and aggressive maneuvering scenarios, such as high-G turns, rapid attitude transitions, aerodynamic saturation effects, and sideslip-induced nonlinear behaviors. Therefore, extending the localization framework to explicitly account for these challenging flight dynamics will be an important direction for future research.

Second, while the proposed framework demonstrates strong performance under standard communication and sensing conditions, practical multi-UAV deployments inevitably face severe operational constraints, including GPS outages in GNSS-denied environments, varying sensor signal-to-noise ratios (SNRs), intermittent packet losses, communication delays, and dynamic network topology disconnections. A comprehensive sensitivity analysis under these degraded sensing and communication conditions has not yet been conducted and remains a limitation of the current study. Future work will therefore focus on systematically evaluating and improving the robustness of the proposed framework under such adverse conditions.

Finally, to alleviate the inherent black-box nature of deep learning models, future research will investigate hybrid localization architectures that combine data-driven temporal learning with model-based filtering and estimation methods. Such integration is expected to further enhance interpretability, robustness, and generalization capability while maintaining real-time execution on resource-constrained onboard platforms.

Author Contributions: Conceptualization, Q.L. and H.L.; methodology, Q.L. and Y.O.; software, Q.L. and Y.O.; validation, Q.L. and Y.O.; formal analysis, Q.L.; investigation, Q.L. and Y.O.; resources, H.L.; data curation, Q.L.; writing—original draft preparation, Q.L.; writing—review and editing, Y.O. and H.L.; visualization, Q.L. and Y.O.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the grants from Science & Technology Department of Sichuan Province of China Nos.2021ZDZX0006 and in part by the Aircraft Swarm Intelligent Sensing

and Cooperative Control Key Laboratory of Sichuan Province, and in part by AVIC United Technology Center for Intelligent Decision-making and Coordinated Control Mechanism Model Research.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Dai, B.; He, Y.; Yang, L.; Su, Y.; Yue, Y.; Xu, W. SIMSF: A scale insensitive multi-sensor fusion framework for unmanned aerial vehicles based on graph optimization. *IEEE Access* **2020**, *8*, 118273–118284. [[CrossRef](#)]
2. Ye, X.; Song, F.; Zhang, Z.; Zeng, Q. A review of small UAV navigation system based on multisource sensor fusion. *IEEE Sens. J.* **2023**, *23*, 18926–18948. [[CrossRef](#)]
3. Geneva, P.; Eckenhoff, K.; Lee, W.; Yang, Y.; Huang, G. Openvins: A research platform for visual-inertial estimation. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: Piscataway, NJ, USA, 2020; pp. 4666–4672.
4. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.; Tardós, J.D. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890.
5. Yue, K. Multi-sensor data fusion for autonomous flight of unmanned aerial vehicles in complex flight environments. *Drone Syst. Appl.* **2024**, *12*, 1–12. [[CrossRef](#)]
6. Zhuang, Y.; Sun, X.; Li, Y.; Huai, J.; Hua, L.; Yang, X.; Cao, X.; Zhang, P.; Cao, Y.; Qi, L.; et al. Multi-sensor integrated navigation/positioning systems using data fusion: From analytics-based to learning-based approaches. *Inf. Fusion* **2023**, *95*, 62–90.
7. Almassri, A.M.; Shirasawa, N.; Purev, A.; Uehara, K.; Oshiumi, W.; Mishima, S.; Wagatsuma, H. Artificial neural network approach to guarantee the positioning accuracy of moving robots by using the integration of IMU/UWB with motion capture system data fusion. *Sensors* **2022**, *22*, 5737. [[CrossRef](#)] [[PubMed](#)]
8. Yeong, D.J.; Panduru, K.; Walsh, J. Exploring the Unseen: A Survey of Multi-Sensor Fusion and the Role of Explainable AI (XAI) in Autonomous Vehicles. *Sensors* **2025**, *25*, 856. [[CrossRef](#)] [[PubMed](#)]
9. Gao, L.; Battistelli, G.; Chisci, L. Fusion of labeled RFS densities with minimum information loss. *IEEE Trans. Signal Process.* **2020**, *68*, 5855–5868. [[CrossRef](#)]
10. Wang, D.; Xu, X.; Yao, Y.; Zhang, T. Virtual DVL reconstruction method for an integrated navigation system based on DS-LSSVM algorithm. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 8501913. [[CrossRef](#)]
11. Tian, B.; Zuo, Z.; Wang, H. Leader–follower fixed-time consensus of multi-agent systems with high-order integrator dynamics. *Int. J. Control* **2017**, *90*, 1420–1427. [[CrossRef](#)]
12. Ning, B.; Han, Q.L. Prescribed Finite-Time Consensus Tracking for Multiagent Systems With Nonholonomic Chained-Form Dynamics. *IEEE Trans. Autom. Control* **2019**, *64*, 1686–1693. [[CrossRef](#)]
13. Wuchengfeng, J.; Jin, C.; Guo, X.; Cheng, X.; Hu, E.; Hang, Q. Development and Prospects of Cooperative Localization and Navigation Countermeasure Technologies for Aircraft Clusters. *J. Astronaut.* **2022**, *43*, 138–142.
14. Hu, J.; Tian, J.; Zou, S.; Chao, T.; Yang, M. A UAV Positioning Algorithm Combining RSS/TOA/INS Based on Adaptive Kalman Filtering. *Unmanned Syst. Technol.* **2022**.
15. Liu, Y.X. Research on Key Technologies of Multi-source Fusion Localization for Unmanned Aerial Vehicle Swarms. Ph.D. Thesis, Beijing University of Posts and Telecommunications, Beijing, China, 2021.
16. She, F.J. Research on Cooperative Location Methods of Multiple Unmanned Aerial Vehicles in GPS-Denied Environments. Master's Thesis, Academy of Military Sciences, Beijing, China, 2020.
17. Chen, H.; Wang, X.-B.; Liu, J.; Wang, J.; Ye, W. Collaborative multiple UAVs navigation with GPS/INS/UWB jammers using sigma point belief propagation. *IEEE Access* **2020**, *8*, 193695–193707.
18. Shule, W.; Almansa, C.M.; Queralta, J.P.; Zou, Z.; Westerlund, T. UWB-based localization for multi-UAV systems and collaborative heterogeneous multi-robot systems. *Procedia Comput. Sci.* **2020**, *175*, 357–364.
19. Scheiber, M.; Cardaillac, A.; Brommer, C.; Weiss, S.; Ludvigsen, M. Modular multi-sensor fusion for underwater localization for autonomous ROV operations. In *Proceedings of the OCEANS 2022, Hampton Roads*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
20. Brommer, C.; Jung, R.; Steinbrener, J.; Weiss, S. MaRS: A modular and robust sensor-fusion framework. *IEEE Robot. Autom. Lett.* **2020**, *6*, 359–366.
21. Zhu, F.; Ren, Y.; Kong, F.; Wu, H.; Liang, S.; Chen, N.; Xu, W.; Zhang, F. Swarm-lío: Decentralized swarm lidar-inertial odometry. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: Piscataway, NJ, USA, 2023; pp. 3254–3260.
22. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.

23. Narkhede, P.; Walambe, R.; Poddar, S.; Kotecha, K. Incremental learning of LSTM framework for sensor fusion in attitude estimation. *PeerJ Comput. Sci.* **2021**, *7*, e662. [[CrossRef](#)] [[PubMed](#)]
24. Abbasimehr, H.; Paki, R. Improving time series forecasting using LSTM and attention models. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 673–691.
25. Wen, X.; Li, W. Time series prediction based on LSTM-attention-LSTM model. *IEEE Access* **2023**, *11*, 48322–48331.
26. Sun, J.; Guo, W. Time series prediction based on time attention mechanism and lstm neural network. In *Proceedings of the 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
27. Kim, Y.; Yoon, S.; Kim, S.; Kim, A. Unsupervised balanced covariance learning for visual-inertial sensor fusion. *IEEE Robot. Autom. Lett.* **2021**, *6*, 819–826. [[CrossRef](#)]
28. Xu, Z.; Zhu, F.; Zhang, X. State Estimation in Multi-Sensor Fusion Navigation: Equivalence Analysis on Filtering and Optimization. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *48*, 1161–1168. [[CrossRef](#)]
29. Zhu, F.; Xu, Z.; Zhang, X.; Zhang, Y.; Chen, W.; Zhang, X. On state estimation in multi-sensor fusion navigation: Optimization and filtering. *arXiv* **2024**, arXiv:2401.05836.
30. Reuter, S.; Vo, B.T.; Vo, B.N.; Dietmayer, K. The labeled multi-Bernoulli filter. *IEEE Trans. Signal Process.* **2014**, *62*, 3246–3260. [[CrossRef](#)]
31. Vo, B.N.; Vo, B.T.; Hoang, H.G. An efficient implementation of the generalized labeled multi-Bernoulli filter. *IEEE Trans. Signal Process.* **2016**, *65*, 1975–1987. [[CrossRef](#)]
32. Irfan, M.; Dalai, S.; Trslic, P.; Riordan, J.; Dooly, G. LSAF-LSTM-Based Self-Adaptive Multi-Sensor Fusion for Robust UAV State Estimation in Challenging Environments. *Machines* **2025**, *13*, 130.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
34. Ma, X.; Luan, H.; Zhao, S.; Niu, T.; Liu, X. Prediction of rolling bearing performance degradation degree based on SELSTM. In *Proceedings of the 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*; IEEE: Piscataway, NJ, USA, 2022; pp. 716–720.
35. Lyu, C.; Yang, B.; Tian, J.; Jin, J.; Ge, C.; Yang, J. Three-fingers FBG tactile sensing system based on squeeze-and-excitation LSTM for object classification. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 7004611.
36. Xiao, K.; Ma, L.; Tan, S.; Cong, Y.; Wang, X. Implementation of uav coordination based on a hierarchical multi-uav simulation platform. In *Proceedings of the Advances in Guidance, Navigation and Control: Proceedings of 2020 International Conference on Guidance, Navigation and Control, ICGNC 2020, Tianjin, China, 23–25 October 2020*; Springer: Singapore, 2022; pp. 5131–5143.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.