applied sciences

MDPI

Article

Long 3D-POT: A Long-Term 3D Drosophila-Tracking Method for Position and Orientation with Self-Attention Weighted Particle Filters

Chengkai Yin [1], Xiang Liu [1,*], Xing Zhang [2], Shuohong Wang [3] and Haifeng Su [4,*]

[1] School of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; m325121528@sues.edu.cn
[2] School of Management, Shanghai University of Engineering Science, Shanghai 201620, China; xzhang@sues.edu.cn
[3] Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA; wangsh@fas.harvard.edu
[4] School of Life Sciences, Shanghai University, Shanghai 200444, China
* Correspondence: xliu@sues.edu.cn (X.L.); hfsu@shu.edu.cn (H.S.); Tel.: +86-13817189293 (X.L.)

Abstract: The study of the intricate flight patterns and behaviors of swarm insects, such as drosophilas, has long been a subject of interest in both the biological and computational realms. Tracking drosophilas is an essential and indispensable method for researching drosophilas' behaviors. Still, it remains a challenging task due to the highly dynamic nature of these drosophilas and their partial occlusion in multi-target environments. To address these challenges, particularly in environments where multiple targets (drosophilas) interact and overlap, we have developed a long-term Trajectory 3D Position and Orientation Tracking Method (Long 3D-POT) that combines deep learning with particle filtering. Our approach employs a detection model based on an improved Mask-RCNN to accurately detect the position and state of drosophilas from frames, even when they are partially occluded. Following detection, improved particle filtering is used to predict and update the motion of the drosophilas. To further enhance accuracy, we have introduced a prediction module based on the self-attention backbone that predicts the drosophila's next state and updates the particles' weights accordingly. Compared with previous methods by Ameni, Cheng, and Wang, our method has demonstrated a higher degree of accuracy and robustness in tracking the long-term trajectories of drosophilas, even those that are partially occluded. Specifically, Ameni employs the Interacting Multiple Model (IMM) combined with the Global Nearest Neighbor (GNN) assignment algorithm, primarily designed for tracking larger, more predictable targets like aircraft, which tends to perform poorly with small, fast-moving objects like drosophilas. The method by Cheng then integrates particle filtering with LSTM networks to predict particle weights, enhancing trajectory prediction under kinetic uncertainties. Wang's approach builds on Cheng's by incorporating an estimation of the orientation of drosophilas in order to refine tracking further. Compared with those methods, our method performs with higher accuracy on detection, which increases by more than 10% on the F1 Score, and tracks more long-term trajectories, showing stability.

Keywords: swarm intelligence; object tracking; deep learning; particle filter; self-attention

Citation: Yin, C.; Liu, X.; Zhang, X.; Wang, S.; Su, H. Long 3D-POT: A Long-Term 3D Drosophila-Tracking Method for Position and Orientation with Self-Attention Weighted Particle Filters. *Appl. Sci.* **2024**, *14*, 6047. https://doi.org/10.3390/app14146047

Academic Editor: Chilukuri K. Mohan

Received: 1 May 2024
Revised: 7 July 2024
Accepted: 8 July 2024
Published: 11 July 2024

1. Introduction

Swarm insects' intricate flight patterns and behaviors have captivated researchers from various disciplines [1–3]. Drosophilas, commonly known as fruit flies, stand out among these insects due to their complex aerial maneuvers and social behaviors. Understanding these behaviors can provide valuable insights into biological phenomena and computational algorithms [4,5]. However, to delve deep into the study of drosophilas' behaviors, we must first be able to track them accurately, especially in 3D environments where they exhibit their full range of motion.

The selection of drosophila as the model organism is driven by its established advantages in scientific research. Its genetic tractability, short lifecycle, and extensive behavioral data facilitate in-depth neurobiological and genetic studies. Drosophila's complex behaviors and well-documented neural circuits make it ideal for high-resolution behavioral analysis. Additionally, a rich body of literature supports the methodologies used in this study, ensuring that our findings are robust and widely applicable in biological contexts.

Tracking 3D drosophilas is a computational challenge and a window into understanding their behaviors holistically. A 3D perspective allows researchers to observe and analyze the depth, altitude, and spatial relationships among drosophilas, which are often missed in 2D tracking. This depth of information can be crucial in studies related to swarm behaviors, mating rituals, and predator–prey interactions among drosophilas.

The challenges of multi-view tracking of the 3D position and orientation of flying swarms mainly result from the large swarm quantity, frequent occlusions, similar appearance, tiny body size, and abrupt motion. Different strategies have been proposed to cope with these difficulties.

By increasing the accuracy and reliability of these tracking techniques, our research refines our understanding of drosophila behavior, contributing to ongoing efforts in ecological monitoring and the advancement of bio-inspired robotics. While our approach draws on established methods, it introduces significant refinements that enhance both the theoretical and practical outcomes, facilitating a more effective integration of biological insights into computational models and engineering solutions.

2. Related Works

Over the years, various algorithms have been proposed to track drosophilas. Traditional methods often relied on color-based or shape-based features to detect and track these insects [6]. While these methods were effective in controlled environments, they struggled in scenarios with dynamic backgrounds, rapid drosophila movements, and occlusions. More recent approaches have incorporated machine-learning techniques, which offer improved accuracy but often at the cost of computational efficiency [7]. For instance, some algorithms leverage convolutional neural networks (CNNs) for detection but may falter when drosophilas are partially occluded or closely clustered [8].

Combined with particle filtering, Wu [9] proposed a method that involved tracking targets in two dimensions in each view and then matching these 2D tracks using linear assignment. However, trajectories sometimes break up due to the target not finding its associated detection in one frame. This led to fragmentation that could not be solved even with a relinking step. To address this issue, Wu [10] relaxed the one-to-one matching constraint, reducing trajectory fragmentation. However, it is essential to note that the effectiveness of these methods is highly dependent on the performance of the detection mechanisms. Another effective strategy for tracking objects is identifying any cross-view associations by matching features to reconstruct 3D observations. Once this is done, a cross-frame association can be established based on the 3D observations. Ardekani [11] has successfully applied this strategy to track several drosophilas. However, the key challenge with this approach is to accurately distinguish each detected object in two dimensions and reconstruct 3D observations. It is difficult to track a large number of objects with small body sizes and similar appearances. Wang [12] applied particle filtering with an LSTM network to predict the kinetic model of each drosophila more accurately. However, the trajectories break when targets are occluded, and the model performs poorly in tracking long-term trajectories.

Our work builds upon some existing methodologies but introduces significant improvements. Recognizing the challenges posed by the dynamic nature of drosophilas and their frequent partial occlusions in multi-target environments, we have developed a novel long-term 3D drosophila tracking method. This method synergistically combines the strengths of deep learning, specifically Mask-RCNN, with the predictive power of particle filtering. Including the Transformer module further refines the tracking process by

predicting the next state of the drosophila, ensuring that our method remains robust even in challenging tracking scenarios.

In the following sections, we explore the intricacies of our approach, emphasizing its benefits over current techniques and showcasing its exceptional ability to track the long-term trajectories of drosophilas. Section 3 outlines the method employed in this study, beginning with using MOG2 for background subtraction, followed by enhanced detection of drosophilas using an improved Mask R-CNN. The method incorporates the estimation of the drosophila's orientation as a critical tracking feature and culminates with a self-attention-enhanced particle-filtering technique. In Section 4, we introduce the experimental setup and process. The results are presented in three aspects in Section 5. Finally, we discuss the results and draw a conclusion.

3. Methods

The general flowchart depicted in Figure 1 outlines the comprehensive steps taken to track drosophilas in a 3D environment using our proposed method. Below, we provide a step-by-step breakdown of the process.

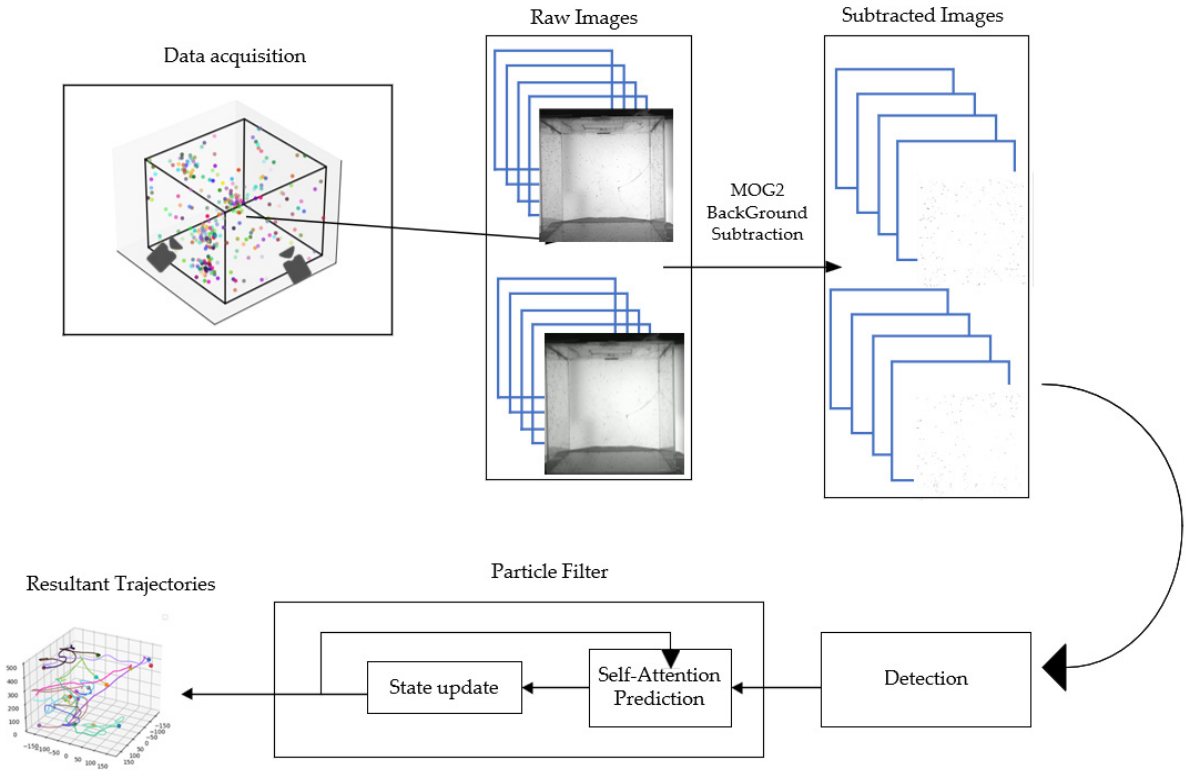


Figure 1. The general flowchart of our method. Each color in the resultant trajectories refer to a trajectory of a drosophila.

Data Acquisition: Utilizing two orthogonally positioned high-speed cameras, videos are captured in a controlled laboratory setting. This dual-camera setup ensures comprehensive spatial coverage, capturing all necessary angles for accurate 3D tracking.

Image Processing: The raw footage obtained from both cameras undergoes background subtraction using the MOG2 algorithm. This step isolates the drosophilas from the static background, focusing only on moving entities within the recorded video.

Detection: Post background subtraction, the detected foreground objects, which primarily include drosophilas, are processed through an improved Mask-RCNN model. This model, which is well-suited for handling partial occlusions, accurately identifies and segments each drosophila by producing segmentation masks. This process not only localizes the drosophilas but also delineates their shapes, which is crucial for the subsequent tracking phase.

Tracking via a Particle Filter: Once the drosophilas are detected, the tracking phase begins using an advanced particle-filtering technique. This particle filter is enhanced with a self-attention mechanism that dynamically updates the weights of the particles. This update is based on the kinematic features of the drosophilas, such as speed and direction, derived from the segmentation masks provided by the Mask-RCNN model.

Trajectory Construction: The final output of the particle-filtering step is the construction of precise 3D trajectories for each tracked drosophila. These trajectories map out the complete movement paths of the drosophilas within the cubic container, illustrating their complex and dynamic flight patterns.

Each of these steps is essential for the effective tracking of drosophilas, providing valuable insights into their behaviors and interactions in a controlled environment.

3.1. Background Subtraction

Background subtraction is a critical step in video processing, especially when tracking objects in dynamic environments. In our methodology, we employ the Mixture of Gaussians 2 (MOG2) algorithm, a widely used approach for robust background subtraction [13]. MOG2 operates on the principle of modeling each pixel as a mixture of Gaussian distributions, which allows for the accommodation of variations in the scene, such as moving backgrounds or changes in illumination.

The MOG2 algorithm updates the background model by employing an online approximation to update the Gaussian mixtures. The probability of observing the current pixel value is given by Equation (1):

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} \cdot \eta(X_t; \mu_{i,t}, \sum_{i,t}) \quad (1)$$

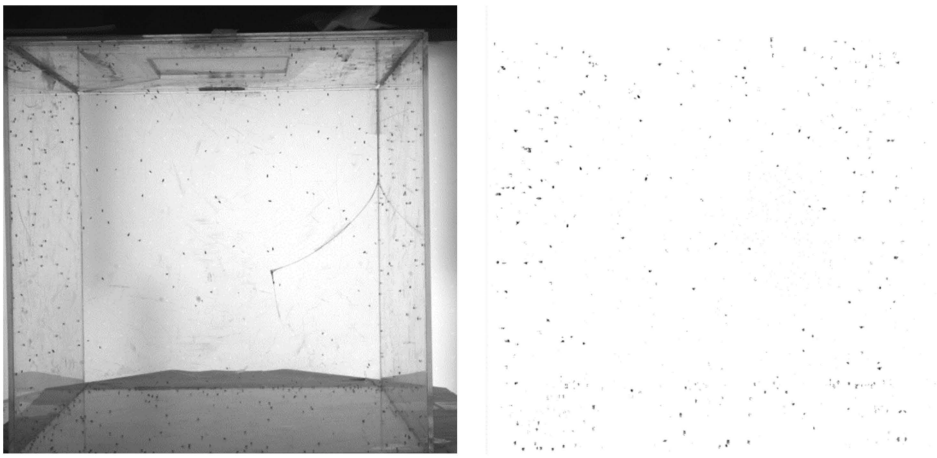
where X_t is the pixel value at time t, K is the number of Gaussian distributions in the mixture, $\omega_{i,t}$ is the weight of the ith Gaussian in the mixture at time t, and η is the Gaussian probability density function with mean $\mu_{i,t}$ and covariance $\sum_{i,t}$.

The algorithm compares the pixel value with the distributions in the model to identify whether a pixel belongs to the foreground or background. If the pixel value does not match the background model well, the algorithm classifies it as part of the foreground. In our case, we are interested in detecting moving drosophilas, which would be considered part of the foreground.

The application of the MOG2 algorithm for background subtraction in our study is driven by specific challenges encountered when tracking drosophila swarms. In the raw video footage, numerous background elements can interfere with accurate tracking, such as the operator's hand during the introduction of drosophilas, the container used for housing them, scratches on the container, and the container's edges. These elements need to be effectively separated from the drosophilas, which are our primary focus.

The MOG2 algorithm is adept at distinguishing moving foreground objects from static or slow-changing backgrounds. By applying this algorithm, we efficiently isolate the drosophilas from irrelevant background elements, which is shown in Figure 2. This separation is crucial, especially considering the small size of individual drosophilas and the potential for numerous background distractions in the video.

In addition, our primary interest lies in tracking the movement of drosophilas. Consequently, drosophilas that remain relatively stationary are considered to be at rest or inactive and are not the focus of our tracking efforts. The MOG2 algorithm aids in this regard by identifying and focusing on moving objects in the video frames. Stationary drosophilas are effectively categorized as part of the background, temporarily suspending their tracking.



(**a**) (**b**)

Figure 2. Visual comparison between an input frame and its subtracted frame. (**a**) Input frame; (**b**) after the subtraction using MOG2.

By removing background noise and distractions, the MOG2 algorithm significantly reduces the likelihood of false detections in the next detection process. This step is particularly important given the small size of the drosophilas and the potential for misidentifying background elements as targets. Additionally, by concentrating on the more active drosophilas, our method aligns with the biological interest in understanding their movement patterns and behaviors.

This approach is particularly effective for our application, as it allows for the detection of drosophilas even in the presence of dynamic changes in the background. The MOG2 algorithm's ability to adapt to changes in lighting and shadowing ensures that the detected foreground consistently represents the drosophilas across different frames.

3.2. Detection

Following background subtraction using the MOG2 algorithm, all detected moving entities are drosophilas, our subjects of interest. While MOG2 effectively identifies these moving targets against the static background, it lacks the precision required for accurate localization and differentiation, especially when drosophilas overlap or are partially occluded.

To enhance the detection accuracy and ensure comprehensive coverage, we utilize a sliding-window technique that segments the video frame into overlapping windows of 128×128 pixels, with a 50% overlap between adjacent windows. As a result, the step size for both horizontal and vertical movement is 64 pixels.

This overlapping design guarantees that drosophilas near the edges of each window are not missed and are adequately captured in multiple windows, thereby improving the chances of their detection and accurate segmentation.

The segmented images are then processed using the Mask-RCNN framework [14], renowned for its precision in instance segmentation. Mask-RCNN builds on the Faster R-CNN architecture by adding a branch for predicting segmentation masks on each Region of Interest (RoI) alongside the existing branches for object classification and bounding-box regression. This setup is particularly effective in handling partial occlusions and densely clustered subjects, as it allows for pixel-precise segmentation.

Through the RoIAlign technique, Mask-RCNN refines the proposed bounding boxes, classifies the objects, and produces detailed segmentation masks that capture the essential features of partially occluded drosophilas. These masks provide crucial details that facilitate the precise determination of each insect's location and orientation.

For the purpose of precise tracking, after each drosophila is detected and segmented in a sliding window, we calculate its global coordinates from its local coordinates within

the window. This is done using the following transformation, where (x_{offset}, y_{offset}) are the coordinates of the top-left corner of the sliding window. The process is given by Equation (2):

$$\begin{aligned} x_{global} &= x_{offset} + x_{local} \\ y_{global} &= y_{offset} + y_{local} \end{aligned} \tag{2}$$

The offsets (x_{offset}, y_{offset}) are calculated based on the indexed position of each sliding window as follows by Equation (3):

$$\begin{aligned} x_{offset} &= j \times S \\ y_{offset} &= i \times S \end{aligned} \tag{3}$$

Here, j and i are the indices of the sliding window in the horizontal and vertical directions, respectively, and S is the step size of 64 pixels. This method ensures the accurate translation of local coordinates detected within each window to the global coordinate system of the entire frame.

The Mask-RCNN model, pre-trained on the COCO dataset [15] and fine-tuned with our annotated frames, outputs bounding boxes, class labels, and segmentation masks for each drosophila detected. These results lay a robust foundation for the particle filter tracking algorithm that follows, ensuring an effective and comprehensive analysis of the complex dynamics and interactions within drosophila swarms. A series of detected drosophila objects is shown in Figure 3, and each cut patch contains one object.

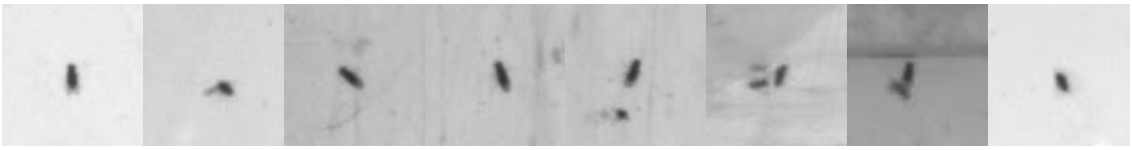


Figure 3. This figure shows a series of drosophila objects detected by our method.

While Mask-RCNN is a potent tool for object detection and segmentation, it faces challenges when tracking small, fast-moving targets like drosophilas, primarily due to the scale variability and its intensive computational demands. We adapted Mask-RCNN by integrating a sliding-window approach and biological modeling of drosophila features to enhance its performance for our specific application. These modifications focus the detection on localized areas and leverage drosophila-specific characteristics, significantly improving the detection accuracy. Additionally, adapting the model through a custom annotated dataset and transfer learning further tailors it to handle the complexities of drosophila tracking in dynamic environments effectively.

3.3. Estimation of Drosophila Orientation

Accurate estimation of the orientation of drosophilas is a critical component of our tracking method. This estimation is particularly challenging due to the small size of the targets and their rapid, erratic movements. Our approach leverages the capabilities of stereoscopic camera views to construct a reliable estimate of each drosophila's orientation in the three-dimensional space.

In each camera view, we model the body of the drosophila as a 2D ellipse. This ellipse fitting is achieved through image segmentation techniques, where the major and minor axes of the ellipse provide a basis for estimating the orientation in each view. The orientation of the ellipse is defined by the angle θ of the major axis for a fixed axis in the image plane. An estimated 2D orientation of a drosophila object is shown in Figure 4.

Figure 4. The estimation of a detected drosophila object. The red line refers to the 2D orientation of the object.

The orientations estimated from each camera view are combined to infer the 3D orientation of the drosophila. This process involves using the known geometrical configuration of the cameras and applying principles of epipolar geometry. The 3D orientation is represented by a vector $\vec{o}$ in the 3D space, which is derived from the 2D orientation angles θ_1 and θ_2 obtained from the two camera views.

The 3D orientation vector $\vec{o}$ is calculated using the following Formula (4):

$$\vec{o} = R \cdot \begin{pmatrix} \cos(\theta_1) \\ \sin(\theta_1) \\ \cos(\theta_2) \end{pmatrix} \tag{4}$$

where R is the rotation matrix derived from the camera calibration process [16].

The estimated 3D orientation of each drosophila is incorporated into the state representation of each particle in the particle filter. This orientation information is crucial as it significantly influences the movement dynamics of the drosophilas, which in turn affects the prediction and update steps of the particle filter.

By accurately estimating the 3D orientation of the drosophilas, our method can more precisely predict their future positions and movements. This accuracy is particularly beneficial in complex scenarios involving rapid changes in direction, occlusions, and interactions among multiple drosophilas.

3.4. Updating Particle Filter Weights with the Kinetics Model

Given the highly nonlinear nature of the dynamic systems of flying swarms like drosophilas and the non-Gaussian distribution of posterior densities, the particle filter framework is employed to approximate the posterior density of the target's state using a set of N weighted particles. Each particle $x_t^{(i)}$ is drawn from the previous state using importance sampling and then adapted according to the dynamic model. The likelihood of each particle given the observed data at time t, denoted $w_t^{(i)}$, scales its contribution. The observation model, $p(z_t | x_t^{(i)})$, plays a critical role in weighting the particles by how likely the observed data are given the particle states. The state estimate of the target at time t is the expected value computed as Equation (5):

$$x_t^E = E(x_t | z_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} x_t^{(i)} \tag{5}$$

The dynamic model in a tracking system predicts the future state of a target based on current observations. In our approach, we integrate a self-attention mechanism within the particle filter's dynamic model, leveraging the deep learning model to learn the kinetic patterns specific to drosophilas' movements. This model accounts for both position and orientation changes dynamically, enhancing the accuracy of the state predictions.

Specifically, we employ a self-attention mechanism to refine the dynamic model by focusing on relevant features of the drosophilas' movement, such as velocity and directional changes, which are crucial for predicting their future state.

Therefore, integrating a self-attention kinetics model significantly enhances the particle filter's weight update mechanism. This integration is pivotal in capturing drosophilas' complex dynamics and interactions in a three-dimensional space, as observed from stereoscopic camera views.

Our decision to employ a self-attention mechanism for predicting the states of particles in the tracking of drosophilas is grounded in empirical studies of their flight behavior. Research indicates that drosophilas exhibit Lévy Flight Trajectories, characterized by a long-tail distribution [17,18]. This pattern typically involves prolonged periods of linear flight interspersed with occasional abrupt maneuvers for evasion or navigation [19]. Such trajectories exhibit a high temporal correlation, where the movement pattern within a certain distance significantly influences the subsequent motion, while the influence of older trajectories diminishes over time.

The flight pattern of drosophilas aligns with the concept of Lévy Flight Trajectories, where the majority of the movement is relatively straightforward, but, occasionally, complex evasive actions are taken. This behavior suggests a strong correlation in the movement patterns over short periods, while longer-term historical data may become less relevant or even misleading for predicting future movements.

Self-attention [20], a key component of the Transformer architecture, allows the model to weigh the importance of different parts of the input sequence. It generates three vectors for each input: Query Q, Key K, and Value V. These vectors are derived through trainable linear transformations of the input data. It allows our method to focus on the most relevant states of the trajectory, considering the influence of recent movements while avoiding the pitfalls of over-emphasizing older historical data.

The core idea of self-attention is to compute a weighted sum of the Value vectors, where the compatibility of the Query with the corresponding Key determines the weights. Mathematically, the attention weights are calculated as in Equation (6):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where d_k is the scaling factor, typically the dimension of the Key vectors. The softmax function ensures that the weights sum up to 1, allowing the model to distribute its focus among the different inputs.

In the context of tracking drosophilas, each particle's state is transformed into Q, K, and V vectors. The self-attention mechanism then dynamically adjusts the weights of the particles based on the learned importance of their features, such as position, velocity, and orientation. This process enables the particle filter to adaptively focus on the most informative particles, thereby improving the accuracy of state estimation.

To align with these behavioral insights, we set the input state length to 20 in our self-attention model. This choice ensures that the model considers a period that is sufficiently informative for predicting the immediate future movements of the drosophilas, without being encumbered by older, less relevant trajectory data. By focusing on this optimal temporal window, the model can more accurately predict the drosophilas' next positions, reflecting their natural movement patterns while minimizing the risk of misinterpretation due to outdated information.

The self-attention model is trained on sequences of drosophila movements, allowing it to capture the complex interaction patterns and movement dynamics. By leveraging these learned patterns, the model can more accurately predict the future states of the drosophilas, which is crucial for updating the particle weights in scenarios involving rapid movement, occlusions, and interactions among multiple targets.

Specifically, our approach is shown step by step in the following pseudocodes (Algorithm 1).

Algorithm 1: self-attention weighted particle filters

Input: Prior distribution $p(x_0)$, system model f, noise q, observation model $p(z_t|x_t^{(i)})$, self-attention model A.
Output: Trajectory $\{\hat{x}_t\}$ for $t = 1$ to T.
// **Initialization**
For $i = 1$ to N do
    Initialize particles $x_0^{(i)}$ from $p(x_0)$
    Set initial weights $w_0^{(i)} = \frac{1}{N}$
End For
For $t = 1$ to T do
    // **Prediction**
    For $i = 1$ to N do
        Predict state $x_t^{(i)} = f(x_{t-1}^{(i)}, q)$
    End For
    // **Update**
    For $i = 1$ to N do
        Calculate unnormalized weight $w_t^{(i)} \propto w_{t-1}^{(i)} \cdot p(z_t|x_t^{(i)})$
    End For
    // **Self-Attention Update**
    For $i = 1$ to N do
        Update weight using self-attention $w_t^{(i)} \propto w_{t-1}^{(i)} \cdot A(x_t^{(i)}, X_t, Z_t)$
        Normalize weights $w_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}$ for all i
        // **Resampling**
        Resample N particles from $\left\{x_t^{(i)}\right\}$ based on weights $\left\{w_t^{(i)}\right\}$
        // **Estimation**
        Calculate state estimate $\hat{x}_t = \sum_{i=1}^{N} w_t^{(i)} \cdot x_t^{(i)}$
    End For
Return: Trajectory $\{\hat{x}_t\}$ for $t = 1$ to T

By aligning our model with the intrinsic characteristics of the drosophilas' movement, we enhance its predictive accuracy and relevance to real-world tracking scenarios.

4. Experiments

Images captured by multiple cameras were applied to test the performance of the proposed method [21–23]. Our method was tested on a reduced number of particles in a controlled environment. We compared trajectory lengths obtained from our experimental data with those from real-world scenarios to validate our tracking system's efficacy. This helped us assess its adaptability in capturing the natural behaviors of drosophilas outside controlled environments. We also evaluated the detection accuracy against manually annotated ground-truth data, which helped us measure our detection algorithm's performance in real-world conditions.

Our experimental setup was meticulously designed to simulate a controlled environment for observing and tracking the flight patterns of drosophilas. The experiments were conducted in a laboratory setting, utilizing a transparent cubic container with an inner dimension of 36 cm to house a specific number of drosophilas. This setup allowed the drosophilas to move freely within a confined space, closely mimicking their natural movement patterns in a relatively open environment.

To ensure the optimal visibility and contrast of the drosophilas against the background, two orthogonal, flat-panel light sources were placed opposite two high-speed monochrome cameras. This arrangement created a backlit illumination setup, significantly enhancing the silhouette of the drosophilas against the transparent container. The cameras were positioned on two sides of the cube, directly facing the light sources, to capture the drosophilas' movements without interference from direct light.

In the experimental design, particular attention was paid to controlling environmental conditions to ensure consistency in the recorded behaviors of drosophilas. The experiments were conducted under standardized laboratory conditions with controlled lighting, temperature, and humidity to minimize variations in drosophila activity that could arise from diurnal or seasonal changes. This setup helped us ensure that the behaviors captured in the video recordings were representative of the drosophilas' natural responses to the experimental setup rather than external environmental factors. By maintaining a consistent environment, we aimed to ensure that the observations reflected inherent behavioral patterns of the drosophilas, which are crucial for developing a reliable tracking system.

4.1. Camera Configuration and Calibration

The cameras were positioned approximately 900 mm away from the container, and their geometric relationship was precisely determined by capturing a standard calibration board (chessboard pattern) using calibration tools. This calibration process ensured accurate spatial alignment and synchronization between the cameras, which are critical to the subsequent 3D trajectory reconstruction. The videos were captured in a cubic transparent acrylic box of size 40 $\times$ 40 $\times$ 40 cm using two geometrically calibrated and temporally synchronized high-speed monochrome CMOS cameras (Manufacturer: IO Industries Inc. Country: Canada, City: London, Ontario, Device name: Flare 4M 180-CL).

To capture the rapid movements of drosophilas, the cameras were set to a frame rate of 100 frames per second. Synchronization between the cameras was achieved through hardware pulses, ensuring that each frame captured by both cameras corresponded to the same instant in time. The resolution of each camera was set to 2048 $\times$ 2048 pixels, and the frame rate was set to 100 FPS, providing the high-detail imagery necessary for the accurate detection and tracking of the small and fast-moving drosophilas.

4.2. Data Collection and Tracking

The experimental procedure involved recording videos of drosophilas' flight from two orthogonal views, utilizing high-speed cameras. This dual-view setup allowed for comprehensive coverage of the flight trajectories, capturing the complex and dynamic movements of the drosophilas within the cubic container. The recorded videos served as the primary data source for our tracking system, which applied advanced computer vision and machine learning techniques, including Mask-RCNN, for detection and segmentation and a self-attention-enhanced particle filter for dynamic state estimation.

To enhance the performance of our Mask-RCNN model, initially pre-trained on the COCO dataset [15], we selected a random subset of 100 frames from the recorded videos for manual annotation. Each frame was meticulously annotated to identify between 150 and 200 drosophila targets along with their corresponding masks, creating a custom dataset tailored to our specific tracking needs. This annotated dataset was used not only for fine-tuning the Mask-RCNN model through transfer learning but also as the ground truth for validating the detection results. Importantly, 10% to 15% of the drosophilas in each frame were partially occluded. The dataset was partitioned into 70% for training, 20% for validation, and 10% for testing, ensuring the rigorous assessment and optimization of the model's performance.

The combination of backlit illumination, precise camera calibration, and high-frame-rate recording was instrumental in achieving high-quality video data. These data were subsequently processed to extract accurate 3D trajectories of the drosophilas, demonstrating the effectiveness of our tracking methodology in capturing the intricate details of their flight patterns. This comprehensive approach, from detailed annotation to targeted training of the detection model, significantly enhanced the model's capability to accurately recognize and segment drosophilas even in challenging scenarios characterized by partial occlusions and rapid movements.

5. Results

5.1. Comparison of Drosophila Detection Methods

Our method demonstrates a significant improvement in detecting the position and state of drosophilas compared with previous approaches. We benchmarked our approach against several methods, including Ameni's method [24], which is designed for larger, more regular targets, Cheng [22] and Wang [12], which both utilize the difference between consecutive frames to detect drosophila positions, and the widely used YoloX model [25]. This comprehensive comparison was conducted using data from our manually annotated dataset, the composition of which is detailed in Section 4.2 of this paper. This comprehensive comparison allowed us to demonstrate the tailored efficiency of our method in the specific context of drosophila tracking, where the nuances of insect motion and occlusion are particularly challenging.

It should be noted that Cheng [22] and Wang [12] actually used the same detection method in their research, which identifies drosophila positions by analyzing differences between consecutive frames. This methodological similarity is particularly evident as Cheng is a co-author with Wang in Wang's paper [12]. For more generalized methods like those of Ameni [24] and YoloX [25], designed primarily for detecting larger and more uniformly moving objects such as aircraft, their performance on tiny, fast, and nonlinear moving targets like drosophila is less effective. In particular, Ameni's approach, tailored for aircraft targets, shows lower accuracy and recall rates when applied to drosophila. Although YoloX exhibits high precision, its recall performance is limited due to inadequate handling of small-scale features. In comparison, our method significantly surpasses these techniques across key metrics such as precision, recall, and F1 Score, demonstrating superior detection capabilities in tracking drosophila, thus highlighting the necessity and effectiveness of our specialized approach.

The comparative analysis focused on three key metrics: precision, recall, and F1 Score. Our method outperformed the other approaches across all these metrics, indicating a higher accuracy in detecting drosophilas. Specifically, the use of the detection method in our approach allowed for more precise segmentation and identification, even in cases of partial occlusion and overlapping flies, which posed challenges for the other methods.

A detailed comparison is presented in Figure 5, which clearly illustrates the superiority of our method in terms of detection performance.

Figure 5. The comparison of detection performance.

5.2. Trajectory Analysis and Comparison

In addition to detection accuracy, we expanded our comparative analysis to include the performance of our method against those of Ameni [24], Cheng [22], and Wang [12] in generating drosophila trajectories. Our evaluation focused particularly on the number and lengths of the trajectories generated. It is important to note that, while Ameni's method is a newer method, it was not specifically designed for tracking extremely small targets like drosophilas, which exhibit unique movement patterns such as Lévy flights. Including Ameni's method in this analysis underscores the fact that general tracking algorithms often underperform when applied to such specialized tracking tasks. This context sets the stage for showcasing the tailored capabilities of our method, which significantly outperforms traditional approaches in tracking the nuanced behaviors of drosophilas.

The histogram in Figure 6 showcases the distribution of trajectory lengths obtained using the different methods. Our approach not only produced a greater number of overall trajectories but also excelled in generating longer trajectories. This indicates a more consistent and reliable tracking performance over time, which is crucial for understanding the movement patterns and behaviors of drosophilas.

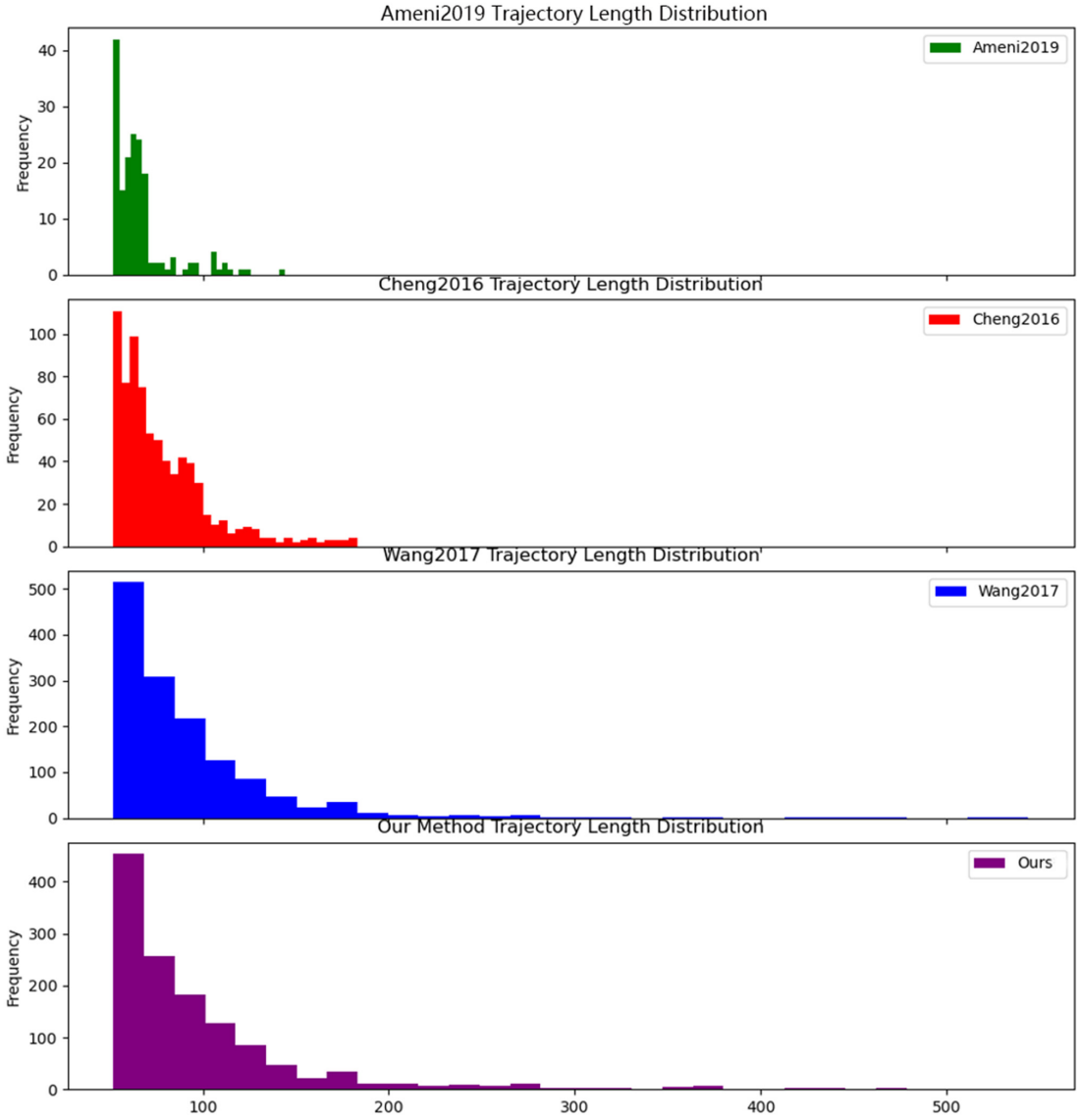


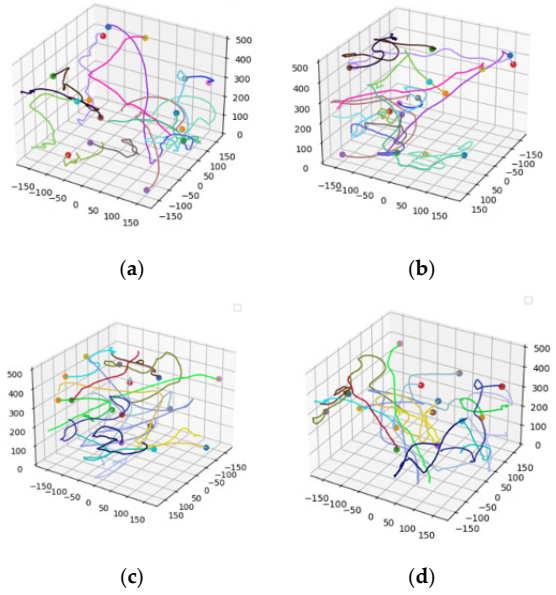
Figure 6. The comparison of the distribution of trajectory lengths obtained using the different methods.

The histogram clearly highlights the shortcomings of Ameni's method when applied to drosophila tracking. While Ameni's method is a newer algorithm, it is primarily optimized for tracking larger objects such as aircraft, which have distinct features and more predictable motion patterns. This specialization renders it less effective for tracking drosophilas, whose small size and random movement patterns (characterized by Lévy flights) result in predominantly short, fragmented trajectories under this method. This issue underlines a general challenge in the field: common tracking methods often fail to perform well with extremely small targets like drosophilas. In contrast, our method demonstrates a significant improvement over both Cheng's method and Wang's method, particularly in sustaining long and consistent trajectories. This improvement is attributed to the robustness of our integrated self-attention particle filter and the accurate orientation estimation from stereoscopic camera views, which effectively handle rapid movements and complex interactions among drosophilas.

5.3. Visualization of Long-Distance Trajectories

A crucial aspect of our results is the visualization of drosophila trajectories in three-dimensional space. To illustrate the effectiveness of our tracking methodology, we selected two sets of 20 long-distance trajectories and visualized them from different perspectives. Each trajectory has more than 200 time steps. These visualizations not only demonstrate the accuracy of our tracking system but also provide insights into the complex movement patterns of drosophilas.

The results of the visualization are shown in Figure 7. Each pair of demonstrated trajectories is tracked from two views of videos.



(**a**) (**b**)

(**c**) (**d**)

Figure 7. The visualizations of two sets of long trajectories. Each color refers to a trajectory of a drosophila. (**a**) This figure presents the first set of 20 long-distance trajectories as viewed from one angle. The trajectories are plotted in a 3D coordinate system, showcasing the intricate paths taken by the drosophilas over an extended period. (**b**) Offering a different perspective, this figure displays the same set of trajectories from another angle. This alternate view further emphasizes the accuracy of our tracking system in capturing the 3D dynamics of fruit fly movement. (**c**) Similar to the first set, we visualized another group of 20 long-distance trajectories. This figure presents these trajectories from one angle, highlighting the consistency and continuity of our tracking results. (**d**) Providing a complementary perspective, this figure shows the second set of trajectories from a different angle. The variation in viewing angles helps us appreciate the three-dimensional nature of the trajectories and the effectiveness of our tracking approach.

These visualizations serve as a testament to the robustness of our method, particularly in maintaining tracking accuracy over longer durations and complex movement patterns. Visualizing the trajectories from different angles offers a comprehensive understanding of the spatial dynamics involved in drosophila movement. The consistency observed across different perspectives reinforces the reliability of our tracking system in capturing the true nature of drosophila trajectories in a three-dimensional environment.

6. Discussion

This study introduces a comprehensive approach to tracking drosophilas in three-dimensional space, tackling the significant challenges posed by their diminutive size, swift movements, and intricate interaction patterns. By integrating advanced computer vision techniques and machine learning models, particularly an improved Mask-RCNN for precise detection and segmentation, alongside a self-attention-enhanced particle filter for dynamic state estimation, our methodology represents a substantial advancement over traditional techniques.

The enhanced detection capabilities of our modified Mask-RCNN model have resulted in significantly higher precision and recall rates for identifying drosophilas, outstripping traditional methods. This is evidenced by an improvement in F1 Scores, which indicates a superior balance of precision and recall compared with earlier approaches. Moreover, our method consistently generates longer and more reliable trajectories, which is critical in studies where continuous monitoring and detailed behavioral analysis are required.

A key innovation in our approach is the incorporation of a self-attention mechanism within the particle-filtering process. This adaptation allows the model to selectively concentrate on the most informative features dynamically, thereby refining the predictions of drosophila movements. Such a capability is particularly effective in environments where drosophilas interact dynamically and are partially occluded, ensuring that our tracking process remains robust even in challenging conditions.

Our approach not only enhances the field of insect tracking but also introduces novel elements like the use of stereoscopic camera views for accurate 3D orientation estimation and the application of self-attention mechanisms in particle filtering. These advancements contribute significantly to a deeper and more nuanced understanding of drosophila dynamics, facilitating more accurate and comprehensive behavioral analyses. The implications of these innovations extend beyond biological research, offering potential enhancements to the design and operation of bio-inspired robotic systems.

Despite the successes of this methodology, there is room for further refinement. An exciting direction for future work involves the development of an end-to-end architecture. By processing raw video data directly to produce trajectories, such a system could achieve higher efficiency and accuracy. This streamlined approach would reduce the complexity and computational demands associated with the current multi-stage processing pipeline, which involves separate stages for detection, orientation estimation, and tracking.

Looking forward, our goal is to not only refine this end-to-end system but also to expand our methodology to include tracking a broader array of insect species across varied environmental settings. Extending the application of our methods could further validate their versatility and robustness, demonstrating their broad utility in ecological monitoring and the study of natural insect behaviors. This would also help in simulating real-world conditions more effectively, where insects often display behaviors influenced by complex environmental interactions. By advancing these capabilities, our method stands to contribute significantly to the fields of ethology, ecology, and robotics, where understanding precise animal movements and behaviors is crucial.

7. Conclusions

In conclusion, the Long 3D-POT method significantly enhances our capability to study intricate insect behaviors within ecological monitoring frameworks. By facilitating precise tracking of drosophilas and other similar entities, this method allows for a comprehensive

analysis of population dynamics, interactions, and environmental responses, which is critical for conservation efforts and pest management strategies.

This sophisticated tracking technology not only serves biological studies but also provides foundational insights necessary for advancing bio-inspired robotics. Understanding the complex movements and behavioral strategies of drosophilas equips us with valuable data to mimic biological efficiency in robot navigation and maneuverability in cluttered environments.

Thus, the advancements introduced by the Long 3D-POT method catalyze significant developments across diverse fields. By improving how we observe and interpret the natural behaviors of tiny, fast-moving insects, we pave the way for innovations that span ecological science and robotic design, enhancing both scientific discovery and practical applications.

Author Contributions: Methodology, C.Y.; Software, S.W.; Validation, X.Z.; Investigation, X.Z. and S.W.; Resources, H.S.; Writing—review & editing, S.W.; Visualization, X.Z.; Project administration, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China University Industry-University-Research Innovation Fund (2021FNB02001).

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: This study did not involve humans.

Data Availability Statement: The data presented in this study is available in this paper.

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

References

1. Theraulaz, G.; Bonabeau, E. A Brief History of Stigmergy. *Artif. Life* **1999**, *5*, 97–116. [CrossRef] [PubMed]
2. Garnier, S.; Gautrais, J.; Theraulaz, G. The biological principles of swarm intelligence. *Swarm Intell.* **2007**, *1*, 3–31. [CrossRef]
3. Fevrier, V. Swarm intelligence: A review of optimization algorithms based on animal behavior. In *Recent Advances of Hybrid Intelligent Systems Based on Soft Computing*; Springer: Cham, Germany, 2021; pp. 273–298.
4. Dankert, H.; Wang, L.; Hoopfer, E.D.; Anderson, D.J.; Perona, P. Automated monitoring and analysis of social behavior in Drosophila. *Nat. Methods* **2009**, *6*, 297–303. [CrossRef] [PubMed]
5. Ramdya, P.; Schneider, J.; Levine, J.D. The neurogenetics of group behavior in Drosophila melanogaster. *J. Exp. Biol.* **2017**, *220*, 35–41. [CrossRef] [PubMed]
6. Fontaine, E.I.; Zabala, F.; Dickinson, M.H.; Burdick, J.W. Wing and body motion during flight initiation in Drosophila revealed by automated visual tracking. *J. Exp. Biol.* **2009**, *212*, 1307–1323. [CrossRef] [PubMed]
7. Kain, J.; Stokes, C.; Gaudry, Q.; Song, X.; Foley, J.; Wilson, R.; de Bivort, B. Leg-tracking and automated behavioural classification in Drosophila. *Nat. Commun.* **2013**, *4*, 1910. [CrossRef] [PubMed]
8. Jiang, Z.; Chazot, P.L.; Celebi, M.E.; Crookes, D.; Jiang, R. Social Behavioral Phenotyping of Drosophila with a 2D–3D Hybrid CNN Framework. *IEEE Access* **2019**, *7*, 67972–67982. [CrossRef]
9. Wu, H.S.; Zhao, Q.; Zou, D.; Chen, Y.Q. Automated 3D trajectory measuring of large numbers of moving particles. *Opt. Express* **2011**, *19*, 7646–7663. [CrossRef] [PubMed]
10. Wu, Z.; Hristov, N.I.; Hedrick, T.L.; Kunz, T.H.; Betke, M. Tracking a large number of objects from multiple views. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009.
11. Ardekani, R.; Biyani, A.; Dalton, J.E.; Saltz, J.B.; Arbeitman, M.N.; Tower, J.; Nuzhdin, S.; Tavaré, S. Three-dimensional tracking and behaviour monitoring of multiple fruit flies. *J. R. Soc. Interface* **2013**, *10*, 20120547. [CrossRef] [PubMed]
12. Wang, S.H.; Su, H.F.; Cheng, X.E.; Liu, Y.; Quo, A.; Chen, Y.Q. Tracking the 3D position and orientation of flying swarms with learned kinematic pattern using LSTM network. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017.
13. Zivkovic, Z. Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; Volume 2.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
15. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13. Springer International Publishing: Berlin/Heidelberg, Germany, 2014.
16. Wang, Y.M.; Li, Y.; Zheng, J.B. A camera calibration technique based on OpenCV. In Proceedings of the 3rd International Conference on Information Sciences and Interaction Sciences, Chengdu, China, 23–25 June 2010.

17. Cole, B.J. Fractal time in animal behaviour: The movement activity of Drosophila. *Anim. Behav.* **1995**, *50*, 1317–1324. [CrossRef]
18. Costa, T.; Boccignone, G.; Cauda, F.; Ferraro, M. The Foraging Brain: Evidence of Lévy Dynamics in Brain Networks. *PLoS ONE* **2016**, *11*, e0161702. [CrossRef] [PubMed]
19. Dubkov, A.A.; Spagnolo, B.; Uchaikin, V.V. Lévy flight superdiffusion: An introduction. *Int. J. Bifurc. Chaos* **2008**, *18*, 2649–2672. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
21. Liu, Y.; Li, H.; Chen, Y.Q. Automatic tracking of a large number of moving targets in 3d. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part IV 12. Springer: Berlin/Heidelberg, Germany, 2012.
22. Cheng, X.E.; Wang, S.H.; Chen, Y.Q. 3D tracking targets via kinematic model weighted particle filter. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016. [CrossRef]
23. Cheng, X.E.; Wang, S.H.; Chen, Y.Q. Estimating orientation in tracking individuals of flying swarms. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.
24. Ellouze, A.; Ksantini, M.; Delmotte, F.; Karray, M. Multiple object tracking: Case of aircraft detection and tracking. In Proceedings of the 2019 IEEE 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey, 21–24 March 2019.
25. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.