Multi-view 6D Pose Estimation of the Aerial Docking Device for Long-Term Drone Operation in Dynamic Environments

Kaixiao Ye[1], Ruoxin Jiang[1], Bize Zhou[1], Weiyu Shao[1], Yuhang Zheng[1], and Tao Yang[1]*

***Abstract*— In aerial refueling and airborne recovery scenarios, accurate localization and 6D pose estimation of the aerial docking device are essential for safe rendezvous and sustained flight coordination. In this paper, we propose a multi-view 6D pose estimation framework for the aerial docking device with distributed observer drones. First, we extract 2D keypoints from multiple views and reconstruct the 3D translation through triangulation and reprojection refinement. Then, we fuse per-view rotation hypotheses with a Gaussian mixture model to resolve symmetry-induced ambiguity. To improve temporal consistency during long-term operation, we close the perception-control loop with a model predictive controller that actively maintains stable observer-target geometry. This integrated design improves observability and robustness of pose estimation during continuous following in dynamic scenarios. Extensive simulation and real-world experiments demonstrate consistent gains in pose accuracy and tracking stability, which supports long-term autonomous drone operation in aerial refueling-related missions. The source code will be publicly available at: `https://github.com/npu-ius-lab/PoseDDF`**

I. Introduction

In aerial refueling and airborne recovery missions, long-term autonomous drone operation depends on reliable perception of the aerial docking device under continuous motion. Because recovery is feasible only within limited time windows along the carrier aircraft trajectory, observer drones must persistently follow the carrier and maintain accurate 6D pose estimation over long horizons. This is challenging in dynamic environments due to viewpoint changes and depth-scale ambiguity in single-view observations. Compared with a single drone, multi-drone multi-view perception provides different viewpoints that improve spatial coverage, observability, and robustness for persistent docking-device perception.

However, multi-view 6D pose estimation and persistent following of the aerial docking device remains challenging, and two fundamental issues must first be addressed. First, due to the geometric symmetry of the aerial docking device, single-view pose estimation is often ambiguous, which leads to inconsistent rotation hypotheses across views [1]. Second, during long-horizon following, observer drones must continuously maintain informative viewpoints and valid triangulation geometry; otherwise, limited FoV and unfavorable relative motion can quickly degrade estimation stability [2].

[1]Kaixiao Ye, Ruoxin Jiang, Bize Zhou, Weiyu Shao, Yuhang Zheng, and Tao Yang are with Unmanned System Research Institute, National Key Laboratory of Unmanned Aerial Vehicle Technology, Integrated Research and Development Platform of Unmanned Aerial Vehicle Technology, Northwestern Polytechnical University, 710072 Xi'an, China. *Corresponding author: yangtao@nwpu.edu.cn

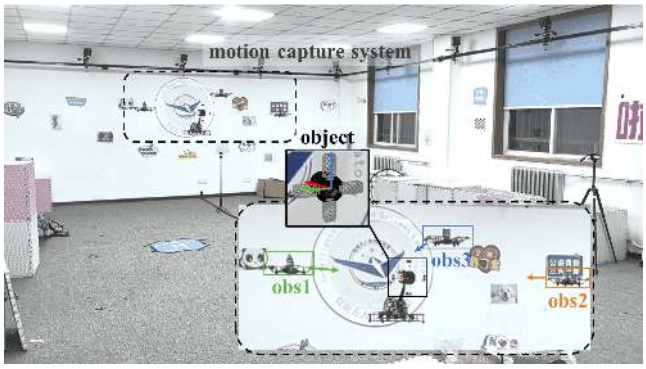


Fig. 1. Visual detection with multiple drones. In our work, we focus on accurate 6D pose estimation and persistent tracking of the aerial docking device, using distributed drones to detect it from different viewpoints.

To address these challenges, this paper proposes a novel method of 6D pose estimation with viewpoints from multiple drones following. Specifically, we first detect target keypoints across multi-drone viewpoints and resolve scale ambiguity through geometric triangulation with the object's 3D model. Then, we employ a Gaussian mixture model (GMM)-based optimization scheme to fuse multi-view estimations, effectively mitigating occlusion and symmetry ambiguity through the complementary observations of viewpoints. Finally, we design a model predictive controller (MPC) to keep the observer drones' following and maintain object 6D pose estimation accuracy. Fig. 1 provides a schematic illustration of our work, where multiple drones collaboratively follow the aerial target to realize 6D pose estimation. The main contributions of this paper are threefold:

- We present a multi-view 6D pose estimation approach for the aerial object with distributed drones, which employs a GMM-based optimization scheme to mitigate the ambiguity of monocular pose estimation of a symmetric target.
- The accuracy and robustness of aerial target 6D pose estimation are improved by the observer drones' continuous following with an MPC controller.
- Both the simulation and real-world experiments demonstrate the effectiveness of our proposed method. The implementation based on ROS nodes will be released to benefit the robotic community.

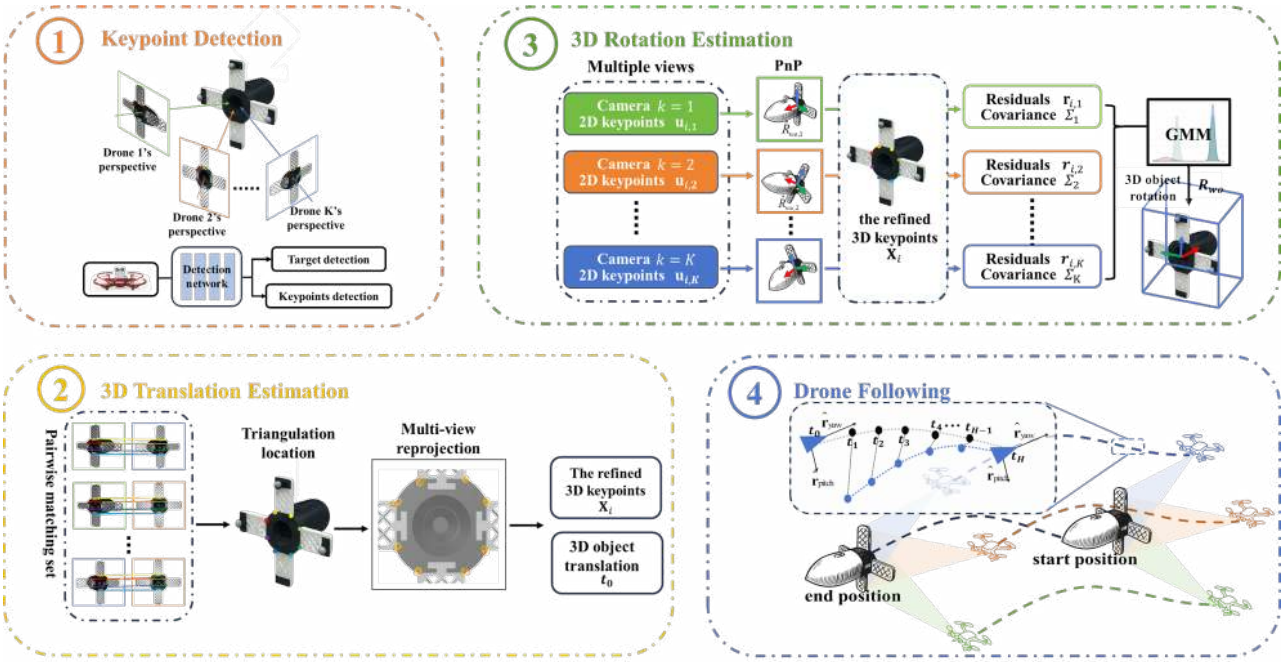


Fig. 2. The pipeline of multi-view 6D pose estimation for the aerial object with distributed drone following.

II. Related Work

A. 6D Pose Estimation

Current research on 6D pose estimation mainly focuses on three levels: instance-level, category-level, and unseen-object-level. Instance-level methods assume a known 3D model and estimate its exact pose [3], [4], category-level methods generalize within object classes [5], [6], and novel-object approaches handle unseen targets via priors [7], [8]. In our scenarios, we focus on instance-level estimation, where multiple collaborative drones estimate 6D pose and keep following the object with a known 3D model. Vision-based 6D pose estimation mainly relies on template matching methods, which align observations with pre-rendered views [9], and feature-based methods, which recover poses from 2D–3D correspondences [10]. While instance-level approaches can achieve high accuracy when the computer-aided design (CAD) models are available, they often struggle with the ambiguity of monocular pose estimation of a symmetric target, motivating the multi-view strategy.

B. Multi-view Detection

Multi-view detection has been widely studied to resolve the scale ambiguity in monocular settings and improve pose estimation accuracy. CosyPose exploited a large number of static views with bundle adjustment to achieve a globally consistent object pose estimation in cluttered scenes [11]. Li et al. [12] proposed a multi-view eye-in-hand framework that fuses keypoint heatmaps into a 3D probability map, enabling robust pose estimation of textureless and reflective objects. Hu et al. [13] developed a dynamic distributed-camera system with gimbal-based adjustments and online extrinsic optimization, allowing robust tracking of moving objects under field-of-view limitations. These methods highlight the benefits of multi-view fusion, but often assume static camera settings.

Gaussian mixture models (GMMs) are widely used in robotics and computer vision to model non-Gaussian noise and multimodal uncertainties. Pfeifer et al. [14] employed an expectation maximization (EM)-based adaptive GMM to dynamically learn noise distributions in factor graph optimization, improving robustness against non-Gaussian sensor errors. Mei et al. [15] leveraged overlap-aware GMM weighting to robustly align partially overlapping point clouds under clutter and noise. Yang et al. [2] introduced a max-mixture formulation of GMM to explicitly address the rotational ambiguity of symmetric objects in multi-view fusion. Inspired by the above methods, we apply GMM to handle the uncertainty of the symmetric object's pose estimation in multi-drone settings.

III. Proposed Method

In this section, we present our framework of multi-view 6D pose estimation for the aerial object with distributed drone following. The proposed method is applicable to any rigid object whose 3D model is available, and we utilize the aerial docking device as the object for 6D pose estimation in this work. First, the object is detected in each view of drones, and the geometric keypoints are extracted simultaneously. Minimizing multi-view reprojection errors, these keypoints are reconstructed into the world coordinate through multi-view triangulation and reprojection refinement. To recover the orientation of a symmetry object, single-view perspective-n-point (PnP) rotations are then fused with GMM to produce a consistent estimate. The estimated poses

are further refined over time using a pose Kalman filter (PoseKF), which improves the accuracy of estimations. Finally, we complete the perception–action loop with an MPC to maintain the object's observability with multiple drones following.

A. Object Keypoint Detection

We adopt YOLOv8 [16] to detect the aerial docking device in each view's image, and extend it with a lightweight keypoint head that predicts N keypoints on the object's edge for view k, $\{\mathbf{u}_{i,k} \in \mathbb{R}^2\}_{i=1}^{N}$. Unlike previous 6D pose estimation pipelines [17] that rely on a centroid and 3D bounding-box corners, we select keypoints on high-contrast geometric boundaries, such as rim vertices and junctions, where image gradients are easily calculated. These keypoints offer two main advantages: (i) Improving cross-view repeatability and better triangulation conditions. (ii) Supporting accurate object rotation estimation by combining multi-view information, thereby mitigating planar-PnP ambiguity and enhancing robustness under partial occlusions.

B. 3D Translation Estimation

Although recent monocular methods can infer depth information from a single image [18], the physical scale is often missing, which can be resolved by multiple views. With known observer drones' poses obtained by internal or external measurements, e.g., GPS, motion-capture system, or visual-inertial odometry, the target's position estimated from each camera view can be transformed into the world coordinate, as illustrated in Fig. 3.

The target's 3D translation is first initialized through a pairwise triangulation of matched feature points from different camera views. We consider that there are M drones. Each has a camera of known pose $T_{wc,k}$ for $k = 1,\ldots,M$. Each camera pose in the world frame is represented as $\mathbf{T}_{wc,k} = [\mathbf{R}_{wc,k}\ \mathbf{t}_{wc,k};\ \mathbf{0}^\top\ 1]$, with camera center $C_k = \mathbf{t}_{wc,k}$. From each detected keypoint $\mathbf{u}_{i,k} = [u,v]^\top$, we compute its corresponding unit bearing vector in the world coordinate via back-projection:

$$\mathbf{d}_{i,k} = \frac{\mathbf{R}_{wc,k}\mathbf{K}^{-1}[u\ v\ 1]^\top}{\left\|\mathbf{R}_{wc,k}\mathbf{K}^{-1}[u\ v\ 1]^\top\right\|} \tag{1}$$

where $\mathbf{K}$ denotes the intrinsic calibration matrix of the camera.

When the corresponding keypoints are detected by the cameras of two drones(a,b) respectively, the 3D position of the corresponding point on the object is estimated by triangulation:

$$P_a = \mathbf{I}_3 - \mathbf{d}_{i,a}\mathbf{d}_{i,a}^\top \tag{2a}$$

$$P_b = \mathbf{I}_3 - \mathbf{d}_{i,b}\mathbf{d}_{i,b}^\top \tag{2b}$$

$$\mathbf{X}_i^{(a,b)} = \arg\min_{\mathbf{X}} \left\{\|P_a(\mathbf{X}-C_a)\|^2 + \|P_b(\mathbf{X}-C_b)\|^2\right\} \tag{2c}$$

where P_a and P_b are projection matrices onto the orthogonal complements of the bearing vectors $\mathbf{d}_{i,a}$ and $\mathbf{d}_{i,b}$, respectively.

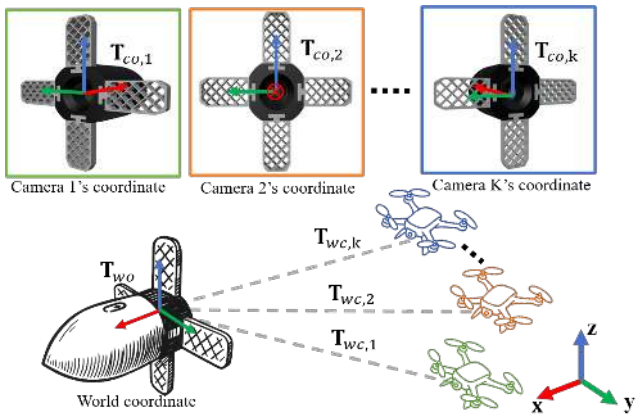


Fig. 3. The transformations from the local camera coordinates to the world coordinate, with known camera extrinsics $T_{wc,k}$.

Therefore, the formulation in (2c) minimizes the squared perpendicular distances from $\mathbf{X}$ to the two rays in different views.

After obtaining the 3D positions of corresponding keypoints, we initialize each keypoint's estimation by averaging the valid pairwise 3D reconstructions. Then, we refine the results through multi-view reprojection optimization. Specifically, each calculated 3D point is back-projected to every camera view and compared to the corresponding keypoint on the image.

$$\min_{\{\mathbf{X}_i\}} \sum_k \sum_i \gamma_{i,k} \left\| \pi\left(T_{cw,k}\,[\mathbf{X}_i^\top\ \ 1]^\top\right) - \mathbf{u}_{i,k} \right\|^2 \tag{3}$$

where $T_{cw,k} = T_{wc,k}^{-1}$, $\gamma_{i,k} \in [0,1]$ denotes a confidence weight. $\pi(\cdot)$ denotes the camera projection function that maps 3D points in the camera coordinate onto the image plane.

Finally, we refine the object's estimated centroid position in the world coordinate as $\mathbf{t}_{wo} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i$, which enables the recovery of a scale-consistent object depth across different drones' views. A reliable 3D translation estimation of the object is essential for subsequent rotation prediction and data fusion.

C. 3D Rotation Estimation

Given the detected 2D keypoints $\{\mathbf{u}_{i,k}\}$ and the known 3D landmarks $\{\mathbf{X}_i^m\}$ in the object model, we first estimate the object's 6D pose in each drone's view by solving a PnP problem:

$$\min_{R_{co,k},\mathbf{t}_{co,k}} \sum_i \left\| \pi\left([R_{co,k}\ \mathbf{t}_{co,k}]\,[\mathbf{X}_i^{m\top}\ \ 1]^\top\right) - \mathbf{u}_{i,k} \right\|^2 \tag{4}$$

PnP returns an estimated pose $(\hat{R}_{co,k}, \hat{\mathbf{t}}_{co,k})$ in the camera coordinate, which is transformed into the world coordinate as

$$\hat{R}_{wo,k} = R_{wc,k}\hat{R}_{co,k} \tag{5}$$

The estimated rotations $\{\hat{R}_{wo,k}\}$ map the model landmarks from the object coordinate into the world coordinate. Then,

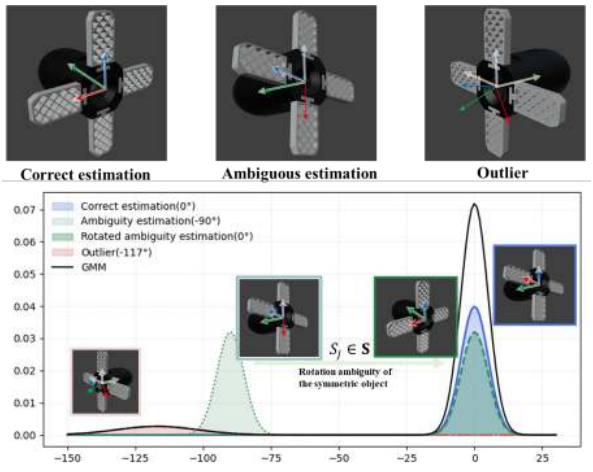


Fig. 4. The top row shows (left to right) correct estimation, ambiguous estimation due to the object's symmetric appearance, and the outlier. The bottom row plots the distribution of rotation hypotheses, where the correct and rectified ambiguous estimations contribute to GMM, and the outlier lies outside the main distribution peak.

these transformed landmarks are compared with the refined 3D keypoints from Sec. III-B. Their deviations are expressed as rotation residuals:

$$\mathbf{r}_{i,k} = \hat{R}_{wo,k}\,\mathbf{X}_i^m - \mathbf{X}_i \tag{6}$$

where $\mathbf{X}_i^m$ represents the model landmarks and $\mathbf{X}_i$ represents the refined 3D keypoints. The rotation covariance in each view is approximated from the Jacobian of the residuals with respect to the rotation parameters:

$$\Sigma_k = \sigma^2 \left(J_k^\top J_k\right)^{-1} \tag{7}$$

where $J_k = \partial \mathbf{r}_{i,k} / \partial \theta_k$ is the Jacobian with θ_k denoting the minimal rotation parameters, and σ^2 is the noise variance of the observations. The defined covariance Σ_k quantifies the reliability of each view's estimation.

To obtain a globally consistent object rotation result R_{wo}, multi-view data fusion is based on a GMM optimization formulation on SO(3):

$$\min_{R_{wo}\in \mathrm{SO}(3)} \sum_k \min_{S_j \in \mathbf{S}} \left\| \mathrm{Log}\left(R_{wo}^\top S_j \hat{R}_{wo,k}\right) \right\|_{\Sigma_k}^2 \tag{8}$$

where $\mathbf{S}$ is the finite set of object symmetries and minimizing over $S_j \in \mathbf{S}$ explicitly resolves symmetry-induced ambiguity by assigning identical costs to physically equivalent rotations. $\mathrm{Log}(\cdot)$ maps a rotation matrix to its tangent vector in $\mathbb{R}^3$. Therefore, all estimations from different drones are fused into a consensus rotation result R_{wo}, and unreliable views due to the object's symmetry are ignored to handle estimation ambiguity. The estimations of rotation R_{wo} and translation $\mathbf{t}_{wo}$ consist of the object's 6D pose in the world coordinate.

$$T_{wo} = \begin{bmatrix} R_{wo} & \mathbf{t}_{wo} \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{9}$$

D. Pose Refinement

We refine the current pose T_{wo} using PoseKF.

The state vector is defined as

$$\mathbf{x} = [x, y, z, \dot{x}, \dot{y}, \dot{z}, \phi, \theta, \psi, \dot{\phi}, \dot{\theta}, \dot{\psi}]^\top \tag{10}$$

which contains position, velocity, attitude (i.e., roll ϕ, pitch θ, yaw ψ), and angular velocities.

The constant velocity model for both translation and rotation is employed as process dynamics. With sampling interval Δt, the state transition is:

$$\begin{aligned} \mathbf{p}_t &= \mathbf{p}_{t-1} + \mathbf{v}_{t-1}\Delta t \\ \mathbf{v}_t &= \mathbf{v}_{t-1} \\ \theta_t &= \theta_{t-1} + \omega_{t-1}\Delta t \\ \omega_t &= \omega_{t-1} \end{aligned} \tag{11}$$

where $\mathbf{p} = [x, y, z]$, $\mathbf{v} = [\dot{x}, \dot{y}, \dot{z}]$, $\theta = [\phi, \theta, \psi]^\top$ and $\omega = [\dot{\phi}, \dot{\theta}, \dot{\psi}]^\top$ the angular velocities.

The corresponding state transition matrix $\mathbf{F}$ is:

$$\mathbf{F} = \begin{bmatrix} \mathbf{I}_3 & \Delta t\,\mathbf{I}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{I}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{I}_3 & \Delta t\,\mathbf{I}_3 \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{I}_3 \end{bmatrix} \tag{12}$$

The measurement vector is defined as:

$$\mathbf{z} = [x, y, z, \phi, \theta, \psi]^\top \tag{13}$$

The measurement matrix is defined as:

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{I}_3 & \mathbf{0}_{3\times3} \end{bmatrix} \tag{14}$$

The process noise covariance $\mathbf{Q}$ and the measurement noise covariance $\mathbf{R}$ are adaptively adjusted based on the residual statistics. For each axis, if the average normalized residual $\bar{r}_i$ exceeds a threshold τ_i, the corresponding measurement noise variance is increased:

$$R_{ii} \leftarrow R_{ii}\,(1 + \alpha(\bar{r}_i - \tau_i)) \tag{15}$$

The process noise is also changed accordingly:

$$Q_{jj} \leftarrow Q_{jj}\,(1 + \beta(\bar{r}_i - \tau_i)) \tag{16}$$

where α and β are adaptation rates. This mechanism balances the reliance between prediction and measurement depending on the reliability of observations.

Finally, the optimized pose is expressed as:

$$T_{wo}^* = \begin{bmatrix} R_{wo}^* & \mathbf{t}_{wo}^* \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{17}$$

E. Drone Following

It's necessary to make the observer drones maintain the formation in front of the target for a stable and accurate object pose estimation. For example, when the target moves out of the camera's FoV, or the observer drones' triangulation baselines degenerate, the estimation becomes unreliable. Therefore, we fix each drone's yaw at a certain angle to the object, and couple perception with control using the MPC. The MPC adjusts the observer drones' positions to preserve a fixed geometric relationship with the object, including a safe observation distance and an optimal line-of-sight angle of $\pi/2$ [19].

We parameterize the control input at each step as translational velocities $\mathbf{u}_\tau = [v_{x,\tau}, v_{y,\tau}, v_{z,\tau}]^\top$ in the drone's body coordinate, with fixed yaw ψ. The control input is transformed into the world coordinate as follows:

$$\begin{bmatrix} v^w_{x,\tau} \\ v^w_{y,\tau} \\ v^w_{z,\tau} \end{bmatrix} = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_{x,\tau} \\ v_{y,\tau} \\ v_{z,\tau} \end{bmatrix} \tag{18}$$

Given the current drone position $\mathbf{p}_\tau = [x_\tau, y_\tau, z_\tau]^\top$, the predicted position at the next step is acquired as:

$$\hat{\mathbf{p}}_{\tau+1} = \mathbf{p}_\tau + dt \cdot [v^w_{x,\tau}, v^w_{y,\tau}, v^w_{z,\tau}]^\top \tag{19}$$

For each predicted trajectory point $\hat{\mathbf{p}}_{\tau+1}$, the reference trajectory is determined jointly by the current drone position and the goal position. Specifically, the goal position is defined by taking the optimized translation $\mathbf{t}^*_{wo}$ as the target location and incorporating the predefined formation geometry. And the reference trajectory is constructed by uniformly interpolating between the current drone position and the goal position within the prediction horizon H, yielding the sequence of reference points $\mathbf{p}^\star_{\tau+1}$. The position residual is:

$$\mathbf{r}^p_{\tau+1} = \hat{\mathbf{p}}_{\tau+1} - \mathbf{p}^\star_{\tau+1} \tag{20}$$

while the control residual is:

$$\mathbf{r}^u_\tau = \mathbf{u}_\tau \tag{21}$$

The MPC output is obtained by minimizing the accumulated residuals over a horizon H:

$$\min_{\{\mathbf{u}_{t:t+H-1}\}} \sum_{\tau=t}^{t+H-1} \left(\|\mathbf{r}^p_\tau\|^2_{Q_p} + \|\mathbf{r}^u_\tau\|^2_R \right) \tag{22}$$

where Q_p weights the position tracking error to keep the drone close to the reference trajectory, while R penalizes the control effort to avoid large velocity commands and ensure smooth motion.

IV. Experiments

A. Experimental Settings

We use an aerial docking device as the object for 6D pose estimation with the multi-drone following. The experimental scenarios involve the observer drones and the target operating in motion, which creates a dynamic and challenging environment for perception and control tasks. In our settings, the system's real-time following performance, its ability to continuously maintain the target within the view, and the localization accuracy of the aerial docking device are all realistic factors that will be considered and evaluated.

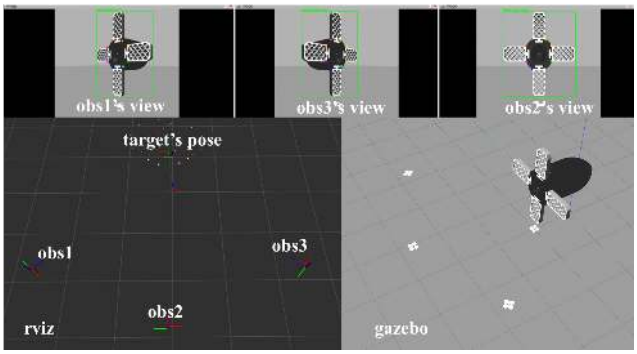


Fig. 5. Screenshots from the simulation experiments.

Our evaluation includes both simulation and real-world experiments, in which multiple observer drones cooperatively track the object and predict its 6D pose. The simulation is implemented on Ubuntu 20.04/ROS Noetic, running on an Intel i9-14900HX processor and 16 GB of RAM. An NVIDIA RTX 4060 GPU is utilized solely for object and keypoint detection. An illustration of the simulation experiments is shown in Fig. 5.

The real-world experiments are conducted using *DJI Tello* quadrotor drones as observers and one custom-built drone carrying an aerial docking device as the target. The ground station is used for image processing, multi-view fusion, and controller output. The *NOKOV* Motion Capture System provides the precise motion states of the observer drones and the ground truth of the target's pose. Due to the coarse extrinsic calibration between the onboard camera coordinate and the observer drone body coordinate, performed with AprilTag [20] and the motion capture system, the initial 6D camera pose contains errors in the centimeter level. During experiments, the observer drones capture image streams and transmit the data to the ground station via WiFi communication. The ground station performs multi-view fusion for object pose estimation and computes the MPC-based control commands. These commands are then transmitted to the observer drones via WiFi, enabling persistent following of the target. All modules are implemented as Robot Operating System (ROS) nodes. An illustration of the real-world experimental environment is shown in Fig. 7.

To quantitatively evaluate the performance of our system, we employ four metrics: the association accuracy (AA), the root mean square error (RMSE), the average distance of object 3D keypoints (ADD/ADD-0.5d), and the 20°20cm criterion.

The AA is to assess the precision of target detection and association across multiple observer drones' views. It is defined as the ratio of correct associated targets to the

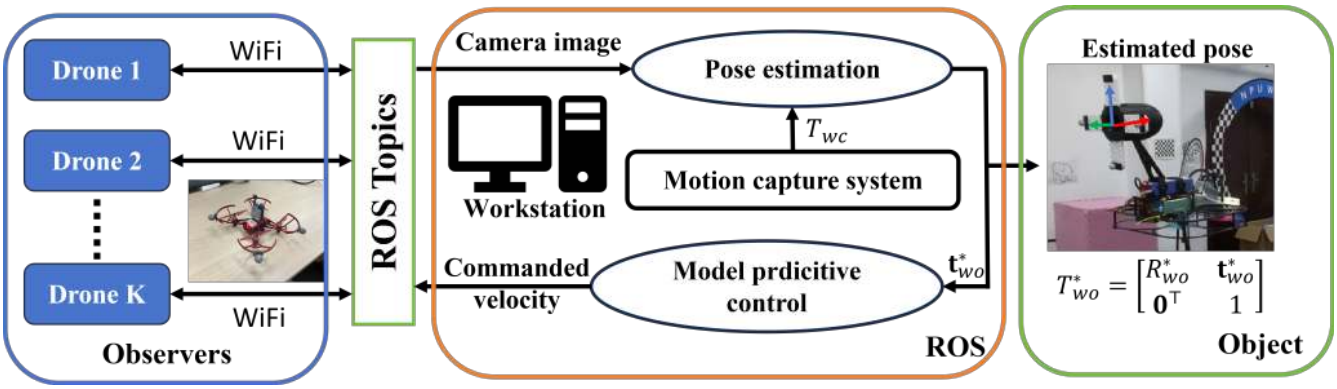


Fig. 6. The real-world experimental settings.

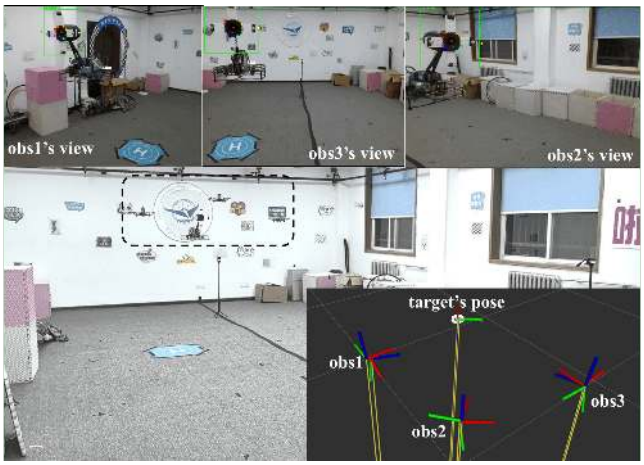


Fig. 7. Screenshots from the real-world experiments.

total number of associations:

$$\mathrm{AA} = \frac{TP}{TP+FP+FN}, \tag{23}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. This metric provides a quantitative measure of how consistently the same target is correctly identified across different viewpoints.

The RMSE evaluates the average translational error between the estimated and ground-truth target positions across time. It is expressed as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| \mathbf{t}_{\mathrm{pred},i} - \mathbf{t}_{\mathrm{gt},i} \right\|_2^2} \tag{24}$$

where $\mathbf{t}_{\mathrm{pred},i}$ is the translation vector predicted in time step i, $\mathbf{t}_{\mathrm{gt},i}$ is the corresponding ground truth of the translation vector , and N is the total number of valid time steps.

The ADD quantifies the discrepancy between the predicted pose and the ground-truth pose by measuring the mean Euclidean distance between the triangulated 3D object keypoints and the real landmarks on the object:

$$\mathrm{ADD} = \frac{1}{|\mathrm{M}|}\sum_{p\in \mathrm{M}} \left\| (R\mathbf{p}+\mathbf{t}) - (\hat{R}\mathbf{p}+\hat{\mathbf{t}}) \right\|_2, \tag{25}$$

where $R, \mathbf{t}$ denote the ground-truth rotation and translation, $\hat{R}, \hat{\mathbf{t}}$ are the estimated ones, and $\mathbf{p}$ is a 3D point in the model set M. In addition, we adopt the ADD-0.5d criterion as a normalized success metric. A prediction is regarded correct if its ADD error is less than half of the object's diameter d, and the final score is reported as the proportion of correct predictions.

The 20°20cm metric jointly evaluates rotation and translation accuracy. The rotation error is defined as:

$$\mathrm{Error}_{rot} = \arccos\left(\frac{\mathrm{trace}(R\hat{R}^\top)-1}{2}\right), \tag{26}$$

and the translation error is

$$\mathrm{Error}_{trans} = \left\| \mathbf{t} - \hat{\mathbf{t}} \right\|_2. \tag{27}$$

If both the rotation error and translation error are below 20° and 20cm, respectively, the estimation is regarded as correct.

B. Experimental Results

1) Simulation Experiments: To evaluate the effectiveness of each proposed module, we conducted comprehensive ablation studies under both two and three observer drones in Table I: (1) Baseline: multi-view PnP estimation followed by simple averaging fusion; (2) Baseline + partial modules: incorporates either PoseKF and MPC to reinforce temporal stability and robust observability, or GMM to effectively disambiguate symmetric rotations; (3) Baseline + GMM + PoseKF + MPC: the full implementation of our method.

The GMM-based fusion module reduces rotation ambiguity effectively, which leads to lower attitude errors and more consistent pose estimation. By handling abrupt motion change or sudden stop of the object, the PoseKF improves temporal stability and reduces RMSE values. The MPC keeps the target near the center of the FoV of each observer drone. This improves the reliability of keypoint detection and object observability, leading to higher AA and localization accuracy. Comparisons between the two-view and three-view configurations also highlight the advantage of leveraging additional viewpoints. The three-view setup consistently achieves lower RMSE and ADD values and higher success rates under the ADD-0.5d and 20°20cm metrics, demonstrating that an additional camera view reduces geometric uncertainty

TABLE I

ABLATION STUDY IN THE SIMULATION.

Views	GMM	PoseKF	MPC	AA ↑	RMSE ↓	ADD ↓	ADD-0.5d ↑	20°20cm ↑
Two views	✗	✗	✗	0.912	0.128	0.213	0.944	0.245
	✗	✓	✓	0.915	0.127	0.165	0.978	0.310
	✗	✗	✓	0.918	0.093	0.163	0.981	0.325
	✓	✓	✓	**0.950**	**0.090**	**0.157**	**0.985**	**0.520**
Three views	✗	✗	✗	0.945	0.110	0.175	0.975	0.505
	✗	✓	✓	0.947	0.108	0.120	0.990	0.630
	✗	✗	✓	0.950	0.072	0.118	0.993	0.648
	✓	✓	✓	**0.980**	**0.070**	**0.115**	**0.996**	**0.884**

and strengthens robustness. When all modules are jointly applied, the full system achieves the best performance in both settings, validating the necessity of combining multi-view perception with robust fusion, temporal filtering, and observer following strategies.

We also ablate the tracking controller by comparing the settings without MPC and with MPC. As shown in Fig. 8, introducing MPC leads to more stable observations and significantly improves both attitude and position estimation accuracy.

2) Real-world Experiments: In the real-world experiments, we validate the effectiveness of the proposed method under both two-view and three-view observer configurations, with the quantitative results summarized in Table II. Compared with the simulation, real-world experiments inevitably suffer from additional sources of uncertainty, such as imperfect self-localization of the observer drones, inaccurate intrinsic and extrinsic camera calibration, and delays or jitter in image transmission.

As a result, the AA with three observer drones is slightly lower than that with two observers. This decline can be attributed to the increased communication load and occasional synchronization issues. Nevertheless, the three-view configuration still demonstrates clear advantages, achieving substantial improvements across other metrics such as RMSE, ADD, and 20°20cm. These results highlight that although AA may decrease slightly due to practical constraints, leveraging multiple views significantly enhances localization accuracy and robustness in real-world pose estimation and persistent drone tracking tasks. Despite inevitable noise and calibration errors in real-world settings, the results exhibit strong consistency with the simulation, thereby demonstrating that the proposed method remains both effective and robust in practical applications.

Table III presents the runtime analysis of each module in our pipeline. The object keypoint detection and translation estimation dominate the overall time consumption. Nevertheless, the full pipeline achieves an average runtime of 58.3 ms, which satisfies real-time requirements. Compared to a comprehensive design, our modular strategy improves the stability of pose estimation efficiently.

TABLE II

PERFORMANCE OF REAL-WORLD EXPERIMENTS.

Metrics	Two views	Three views
AA	**1.000**	0.980
RMSE [m]	0.095	**0.086**
ADD [m]	0.157	**0.116**
ADD-0.5d	0.096	**1.000**
20°20cm	0.391	**0.812**

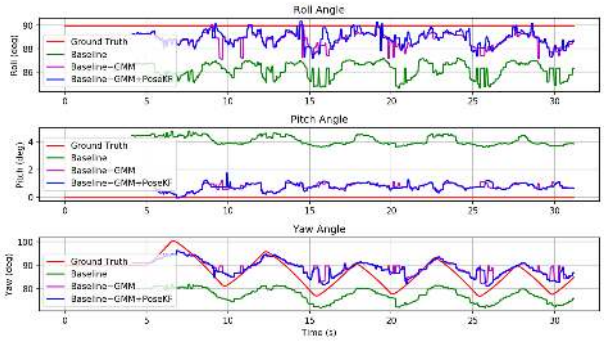
TABLE III

RUNTIME OF EACH MODULE IN THE POSE ESTIMATION PIPELINE.

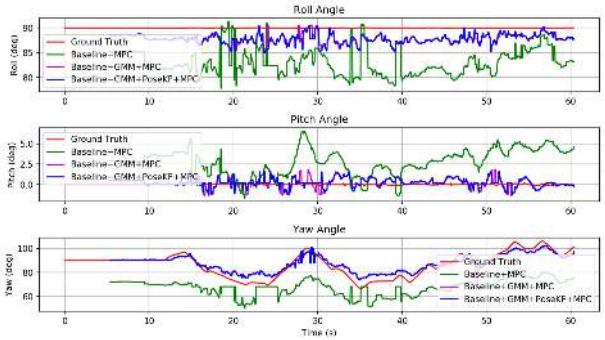
Modules	Runtime (ms)
Object keypoint detection	23.00
3D Translation estimation	20.00
PnP pose regression	2.00
GMM-based rotation fusion	13.00
PoseKF	0.30
Total pipeline	58.3

V. CONCLUSIONS

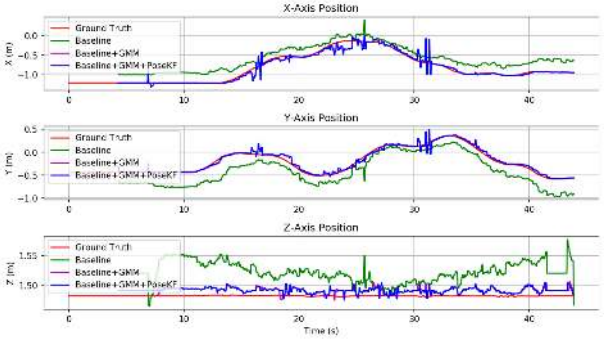
In this work, we presented a framework for persistent tracking and localization of an aerial object using multiple drones. Our method integrated multi-view pose estimation, a PoseKF for temporal refinement, and a model predictive controller, ensuring both accurate 6D pose estimation for the object and stable drone following. By leveraging multi-drone cooperation, the framework effectively mitigated rotation estimation ambiguity of a symmetric object, enhances pose observability, and maintains robustness against viewpoint degeneracy. Comprehensive validation in both simulation and real-world experiments confirmed our system design, demonstrating reliable pose estimation and sustained following of the target in dynamic environments. These results indicate that the proposed framework can maintain stable estimation and following performance during long-horizon continuous operation for the aerial docking device.



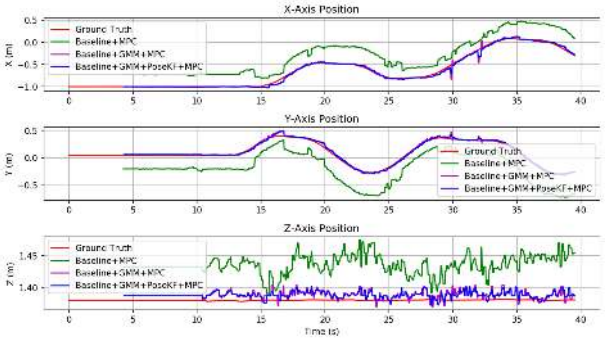
(a) attitude (without drone following)



(b) attitude (with drone following)



(c) Position (without drone following)



(d) Position (with drone following)

Fig. 8. Ablation study evaluating attitude and position estimation.

REFERENCES

[1] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[2] P. Yang and W. Wen, "Tightly joining positioning and control for trustworthy unmanned aerial vehicles based on factor graph optimization in urban transportation," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 3589–3596.

[3] S. Peng, Y. Liu, Q. Huang, H. Bao, and X. Zhou, "Pvnet: Pixel-wise voting network for 6dof pose estimation," 2018. [Online]. Available: https://arxiv.org/abs/1812.11788

[4] Y. Xu, K.-Y. Lin, G. Zhang, X. Wang, and H. Li, "Rnnpose: 6-dof object pose estimation via recurrent correspondence field estimation and pose optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4669–4683, 2024.

[5] X. Deng, J. Geng, T. Bretl, Y. Xiang, and D. Fox, "icaps: Iterative category-level object pose and shape estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1784–1791, 2022.

[6] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield, "Keypoint-based category-level object pose tracking from an rgb sequence with uncertainty estimation," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 1258–1264.

[7] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 868–17 879.

[8] T. Liang, Y. Zeng, J. Xie, and B. Zhou, "Dynamicpose: Real-time and robust 6d object pose tracking for fast-moving cameras and objects," 2025. [Online]. Available: https://arxiv.org/abs/2508.11950

[9] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," 2018. [Online]. Available: https://arxiv.org/abs/1703.10896

[10] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3d orientation learning for 6d object detection from rgb images," 2019. [Online]. Available: https://arxiv.org/abs/1902.01275

[11] Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[12] A. Li and A. P. Schoellig, "Multi-view keypoints for reliable 6d object pose estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6988–6994.

[13] Y. Hu, S. Wang, D. Li, X. Chen, M. Zhu, Z. Xin, and J. Yu, "Adjusting distributed cameras for robust moving object pose estimation," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 10 650–10 659, 2025.

[14] T. Pfeifer and P. Protzel, "Expectation-maximization for adaptive mixture models in graph optimization," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3151–3157.

[15] G. Mei, F. Poiesi, C. Saltori, J. Zhang, E. Ricci, and N. Sebe, "Overlap-guided gaussian mixture models for point cloud registration," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 4500–4509.

[16] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[17] Y. Zheng, C. Zheng, J. Shen, P. Liu, and S. Zhao, "Keypoint-guided efficient pose estimation and domain adaptation for micro aerial vehicles," *IEEE Transactions on Robotics*, vol. 40, pp. 2967–2983, 2024.

[18] Y. Pan, B. Liu, Z. Liu, H. Shen, J. Xu, W. Fu, and T. Yang, "Mona bench: A benchmark for monocular depth estimation in navigation of autonomous unmanned aircraft system," *Drones*, vol. 8, no. 2, p. 66, 2024.

[19] S. Haoran, L. Faxing, W. Hangyu, and X. Junfei, "Optimal observation configuration of uavs based on angle and range measurements and cooperative target tracking in three-dimensional space," *Journal of Systems Engineering and Electronics*, vol. 31, no. 5, pp. 996–1008, 2020.

[20] M. Wheeler, R. Wise, R. Rysdyk, W. Whitacre, and M. Campbell, "Autonomous cooperative geo-location and coordinated tracking of moving targets," in *AIAA Infotech@ Aerospace 2007 Conference and Exhibit*, 2007, p. 2852.