

Article

Robust and Cost-Effective Vision-Based Indoor UAV Localization with RWA-YOLO

Feifei Wang, Kun Sun and Yuanqing Wang *

School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China;
522024230063@smail.nju.edu.cn (F.W.)

* Correspondence: yqwang@nju.edu.cn

Abstract

Accurate indoor localization for unmanned aerial vehicles (UAVs) remains challenging in GPS-denied environments, especially for small-object detection and under low-light conditions. We propose Robust Wavelet-Aware YOLO (RWA-YOLO), a vision-based detection framework that integrates a wavelet-aware attention fusion module with a dual multi-path aggregation mechanism to enhance small-object detection and multi-scale feature representation. UAV-mounted LEDs are utilized to ensure robust visual perception in low-light indoor scenarios. The UAV's three-dimensional position is estimated through multi-view geometric triangulation without relying on external beacons or artificial markers. Beyond static localization, the system is validated under dynamic flight conditions, demonstrating smooth and temporally coherent trajectory reconstruction suitable for real-time control loops (update rate ≈ 25 FPS). Extensive experiments in real indoor environments achieve centimeter-level localization accuracy (root mean square error: 9.9 mm, 95th percentile error: 13.5 mm), outperforming state-of-the-art vision-based methods and achieving accuracy comparable to or better than representative hybrid ultra-wideband–vision systems reported in the literature. These results confirm the effectiveness, robustness, and real-time capability of RWA-YOLO for indoor UAV navigation in constrained environments.

Keywords: indoor localization; UAV positioning; vision-based system; YOLO; illumination robustness

1. Introduction

Unmanned Aerial Vehicles (UAVs) have evolved from specialized military equipment into versatile platforms widely used in civil and industrial domains. Their flexibility, efficient data collection, and reduced operational risks have driven adoption in precision agriculture [1], emergency response [2], environmental monitoring [3], and infrastructure inspection [4]. The increasing demand for accurate indoor positioning has further accelerated the development of UAV localization technologies.

Conventional UAV localization and positioning systems typically rely on Global Navigation Satellite Systems (GNSSs), cameras, inertial sensors, and ranging devices such as LiDAR (Light Detection and Ranging) or ultrasonic sensors. While the GNSS performs well in outdoor environments, its accuracy deteriorates indoors due to severe signal attenuation and multipath effects [5]. This limitation motivates the study of Indoor Positioning Systems (IPSs), which aim to provide reliable and accurate localization in GNSS-denied environments such as warehouses, factories, and public facilities [6].



Academic Editor: Bingbing Gao

Received: 25 December 2025

Revised: 5 February 2026

Accepted: 12 February 2026

Published: 26 February 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.

Indoor positioning remains challenging due to obstacles, furniture, and moving objects, which introduce noise, interference, and non-line-of-sight (NLOS) conditions [7]. The variability of building layouts and materials further complicates the design of universal solutions. UAV applications impose additional requirements, including high precision, low latency, and robustness under environmental variations [8]. Achieving these objectives in real time while maintaining cost-effectiveness is non-trivial.

Accurate indoor positioning is critical for UAV safety and higher-level autonomous functions. Precise localization reduces collision risks and enables advanced applications such as automated warehouse inspection, cooperative multi-UAV missions, and security surveillance in sensitive environments [9]. Existing IPS approaches include ultra-wideband (UWB) [10], ultrasonic systems [4], radio frequency identification (RFID) [11], vision-based methods [12], and hybrid sensor-fusion frameworks [13]. Each approach has demonstrated potential, but limitations such as infrastructure dependence, sensitivity to noise, or high computational demands hinder large-scale deployment.

In this work, we propose a vision-driven indoor positioning method for UAVs that leverages computer vision and triangulation to achieve high-accuracy localization in complex indoor spaces. Unlike RF-based methods that rely on signal strength, our approach utilizes visual features for robust detection and depth estimation, reducing susceptibility to multipath interference. The main contributions of this study are:

1. A cost-effective, infrastructure-free vision-based indoor localization framework for UAVs using distributed cameras.
2. A robust UAV detection model, termed Robust Wavelet-Aware YOLO (RWA-YOLO), where “Robust” emphasizes system-level reliability under challenging indoor conditions, and “Wavelet-Aware” refers to the explicit integration of wavelet-domain feature fusion and multi-path aggregation for enhanced small-object perception.
3. Comprehensive experimental evaluation demonstrating centimeter-level 3D localization accuracy in real indoor environments.

The remainder of this paper is organized as follows. Section 2 reviews the related work on indoor UAV positioning. Section 3 details the hardware architecture and system components. Section 4 introduces the proposed method for UAV detection and localization. Section 5 presents the experimental setup and evaluates the overall system performance. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2. Related Work

Indoor UAV positioning has gained increasing attention due to the limitations of GNSSs in indoor or obstructed environments. Early approaches primarily relied on radio-frequency (RF) techniques, including Wi-Fi, Bluetooth, and RFID. Bluetooth Low Energy (BLE) was evaluated for its indoor localization accuracy [6], while Wi-Fi fingerprinting methods showed susceptibility to multipath interference [7]. Passive RFID-based UAV localization has also been investigated, providing low-cost solutions but with accuracy inferior to UWB systems [11].

UWB technologies have emerged as a leading solution for high-precision indoor UAV positioning. Lin and Zhan [1] combined UWB with visual-inertial odometry, Yang et al. [2] demonstrated UWB-based localization in highly confined spaces, and Tiemann and Wietfeld [3] proposed a scalable time difference of arrival (TDOA)-based multi-UAV localization framework. Recent studies have enhanced UWB performance by integrating additional sensing modalities. For instance, Kao et al. [14] introduced a deep learning-based UWB–vision–IMU fusion framework (VIUNet). Dual-anchor UWB models [15] and hybrid UWB–monocular vision approaches [16] have further improved robustness and localization precision.

Acoustic and ultrasonic positioning methods have also been explored. Ledergerber et al. [4] developed a one-way ultrasonic localization system, which can provide short-range accuracy but is vulnerable to environmental noise and turbulence, limiting large-scale adoption.

Vision-based techniques have gained popularity for infrastructure-free and high-resolution indoor localization. Marker-based methods improve detection reliability under occlusion [8]. SLAM-based frameworks, such as ORB-SLAM [9] and VINS-Mono [17], offer real-time pose estimation in GPS-denied scenarios. Deep learning techniques have been applied for navigation policy learning [12] and low-computation UAV localization [18]. Vision-based autonomous navigation for quadrotor UAVs in unknown and constrained environments has also been investigated, demonstrating the feasibility of perception-driven flight control [19]. In addition, generative-model-based indoor localization methods, such as Wasserstein GAN-based pseudo fingerprint mapping, have been explored for improving localization robustness in complex indoor environments [20]. Recent studies have also explored UWB-based autonomous indoor navigation, highlighting the potential of hybrid RF–vision approaches for robust UAV positioning in GNSS-denied environments [21]. Robustness under varying illumination and occlusion has been addressed in recent works [22,23].

Hybrid sensor-fusion approaches aim to leverage the complementary strengths of multiple modalities. Schmid et al. [13] fused stereo vision with inertial data for accurate UAV navigation. Public datasets, such as MILUV [24], facilitate benchmarking of hybrid algorithms. Integrated vision and RF frameworks, including CIUAV [22], have demonstrated improved accuracy and reliability compared to single-modality systems. Zimmerman et al. [25] combined onboard sensing with floor plans for energy-efficient localization. IVU-AutoNav [16] further demonstrated the effectiveness of integrated visual–UWB navigation.

Visual SLAM and visual odometry (VO) techniques have been widely studied for indoor UAV navigation in GPS-denied environments. These methods typically rely on onboard cameras to estimate the UAV's ego-motion through continuous feature tracking and incremental map construction. While effective for relative pose estimation and trajectory tracking, SLAM/VO approaches often suffer from drift accumulation and require sufficient scene texture and motion continuity.

In contrast, the focus of this work is external vision-based absolute localization using multiple fixed and calibrated cameras. The proposed system directly estimates the UAV's global 3D position in a predefined world coordinate system via multi-view triangulation, without relying on motion continuity or map building. Due to these fundamental differences in sensing configuration, output definition, and evaluation protocol, SLAM/VO methods are not considered as direct baselines in this study.

In summary, RF-based approaches provide high accuracy but often require dense infrastructure. Acoustic methods are lightweight but sensitive to noise. Vision-based solutions are flexible and infrastructure-free but computationally demanding. Hybrid frameworks improve robustness at the cost of added complexity. These observations motivate the vision-driven, cost-effective indoor UAV positioning method proposed in this work.

3. Hardware Architecture and System Components

3.1. Hardware Architecture

To achieve accurate localization of UAVs in GPS-denied environments, this study designs and implements an indoor UAV positioning system. The system adopts a multi-camera configuration integrated with computer vision techniques to perform UAV detection and three-dimensional position estimation. Specifically, four cameras are deployed at different locations within the indoor environment to capture multi-view images. UAV features

are extracted through vision-based recognition algorithms, and the spatial coordinates of the UAV are subsequently computed using multi-view geometry.

As illustrated in Figure 1, the overall hardware framework consists of four camera nodes and a central processing unit. The cameras capture multi-view images of the indoor environment and transmit the image data via a network to the central processing unit. The central processing unit performs UAV detection and localization based on the received camera images. All cameras are connected to the central processing unit via USB interfaces, ensuring efficient image transmission to the detection module for UAV recognition and localization.

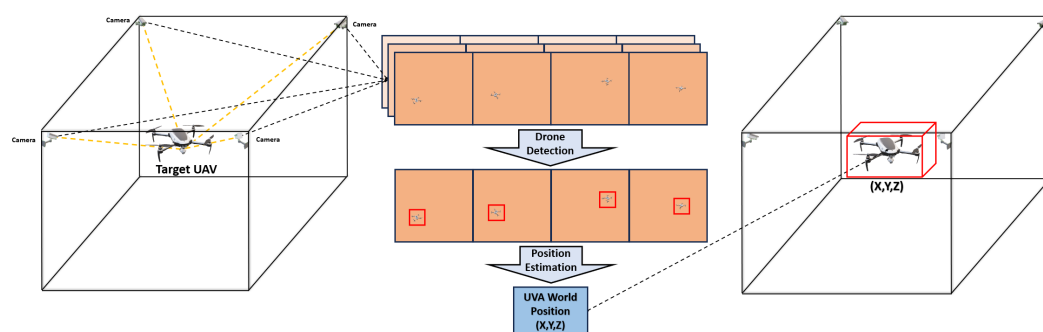


Figure 1. Overall system architecture of the proposed vision-driven UAV indoor localization framework. The system employs four cameras to capture UAV images, applies a UAV detection algorithm to obtain 2D positions in each view, and estimates the UAV's 3D location using a multi-view position estimation method.

3.2. System Components

The system is configured with four low-cost USB cameras, serving as the primary perception units for the indoor UAV localization system. The selected cameras have a resolution of 1920×1080 pixels and a field of view of $96^\circ \times 80^\circ$. They are evenly positioned at the four corners of the indoor space to ensure comprehensive coverage of the experimental area. This distributed multi-camera configuration enables stable image acquisition from multiple viewpoints and supports accurate 3D position estimation through multi-view geometry.

4. Detection and Localization Method

The detection and localization algorithm consists of three core steps:

1. Image acquisition and preprocessing: Multiple cameras synchronously capture indoor scene images.
2. Target detection: The UAV is identified in each camera image using vision-based detection algorithms.
3. 3D position estimation: Based on camera calibration parameters and detection results, multi-view geometry is employed to compute the UAV's spatial position within the indoor coordinate system.

4.1. Image Acquisition and Preprocessing

In the proposed system, four cameras are connected to the central processing unit via USB interfaces. To ensure consistency across multiple viewpoints, a synchronized triggering mechanism is adopted, enabling all cameras to capture indoor scene images with a negligible temporal delay for this particular application. Each camera operates at a resolution of 1920×1080 with a frame rate of approximately 30 fps. The image acquisition process is managed uniformly by the OpenCV library, which ensures stability during data collection.

The acquired raw images require both intrinsic and extrinsic calibration. Intrinsic calibration is performed using a checkerboard pattern to obtain each camera's focal length, principal point coordinates, and distortion parameters, followed by distortion correction to ensure geometric accuracy and to avoid localization errors caused by lens distortion. The intrinsic matrix of the camera is given by (1).

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where f_x, f_y denote the focal lengths along the horizontal and vertical directions, and (c_x, c_y) are the principal point coordinates. The distortion parameters include radial distortion (k_1, k_2, k_3) and tangential distortion (p_1, p_2) , which can be expressed as (2):

$$\begin{aligned} x_d &= x(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_1xy + p_2(r^2 + 2x^2) \\ y_d &= y(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_2xy + p_1(r^2 + 2y^2) \end{aligned} \quad (2)$$

where $r^2 = x^2 + y^2$. By applying this model, distortion can be corrected and geometrically accurate images are obtained (as illustrated in Figure 2 for an example of intrinsic calibration).

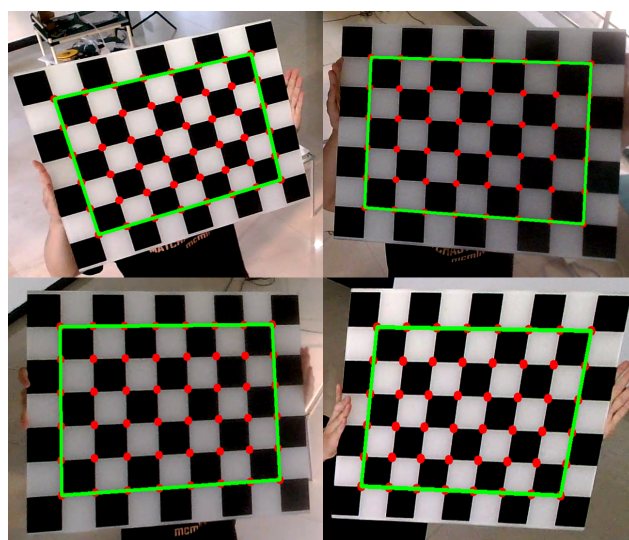


Figure 2. Intrinsic calibration image with detected and labeled checkerboard corners.

Extrinsic calibration is also carried out using a checkerboard, aiming to estimate the rotation matrix R and translation vector T that relate the world coordinate system to the camera coordinate system. The projection relationship can be expressed as (3):

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} R & T \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (3)$$

where (u, v) are image coordinates, (X_w, Y_w, Z_w) are world coordinates, and s is a scale factor. In practice, pixel coordinates are first converted into normalized camera coordinates as (4):

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (4)$$

By combining the known 3D checkerboard corner coordinates $\{x^i, X_w^i\}$, a set of 2D–3D correspondences is established. Consequently, the extrinsic calibration problem can be formulated as the classical Perspective-n-Point (PnP) problem, which is employed exclusively during the offline calibration stage to estimate the fixed camera extrinsic parameters.

$$(R, T) = \text{PnP}\left(\{X_w^i\}, \{x^i\}, K\right) \quad (5)$$

where x^i denotes the corrected image coordinates of the corner points. Since the distance between the camera and the calibration plane is relatively large, automatic corner detection is often unstable. Therefore, a semi-automatic approach is introduced: each camera captures two images of the checkerboard (as illustrated in Figure 3), and 25 corners are manually selected to obtain the initial 2D coordinates. Then, a neighborhood refinement procedure is applied around each manually selected point to reduce human-induced errors, resulting in more reliable estimates of the rotation matrix R and translation vector T .

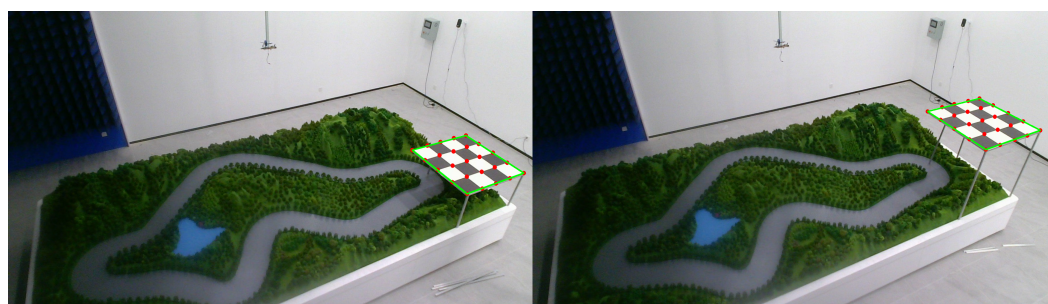


Figure 3. Extrinsic calibration image with manually selected and refined checkerboard corners.

Furthermore, to validate the accuracy of the estimated extrinsic parameters, the reprojection error is calculated, defined as (6):

$$e = \frac{1}{N} \sum_{i=1}^N \|x^i - \hat{x}^i\| \quad (6)$$

where the reprojected point $\hat{x}^i$ is computed using the projection model (7):

$$\hat{x}^i = \pi\left(K(RX_w^i + T)\right) \quad (7)$$

where x^i denotes the actual image coordinates of the corner points, and $\hat{x}^i$ is the reprojected coordinate obtained from the estimated extrinsic parameters. In our system, the reprojection error is used as a quantitative indicator to validate the accuracy of the estimated camera extrinsic parameters. Across all cameras, the average reprojection error is below 0.6 pixels, while the maximum reprojection error does not exceed 1.0 pixel, indicating reliable camera calibration quality for subsequent multi-view 3D position estimation. The experiments were conducted in an indoor room with approximate dimensions of 10 m × 8 m × 5 m (length × width × height), which provides the spatial scale for interpreting the reported pixel-level reprojection error in terms of metric localization accuracy.

4.2. Target Detection

As a mainstream one-stage object detection framework, YOLOv11 demonstrates strong feature extraction capability and excellent generalization performance, achieving remarkable results across a wide range of object detection tasks. Therefore, this study adopts YOLOv11 as the core detection backbone. However, when applied to indoor unmanned aerial vehicle (UAV) detection, the original YOLOv11 architecture still faces several challenges:

1. UAV targets are typically small with limited structural features, which makes them challenging to detect in indoor vision-based localization scenarios;
2. Indoor environments contain rich textures and numerous interferences, which can significantly degrade detection accuracy;
3. UAV detection requires high real-time performance, which constrains the allowable model complexity.

To address the above issues, the YOLOv11 architecture is improved as illustrated in Figure 4. First, the original downsampling modules are replaced with the proposed Wavelet-Aware Fusion (WAF) blocks to more effectively preserve multi-scale feature information. Second, an enhanced Dynamic Multi-scale Patch Attention (DMPA) block is incorporated into the C3k2 module, introducing a wavelet-attention-based fusion mechanism that further improves UAV recognition accuracy under complex indoor conditions.

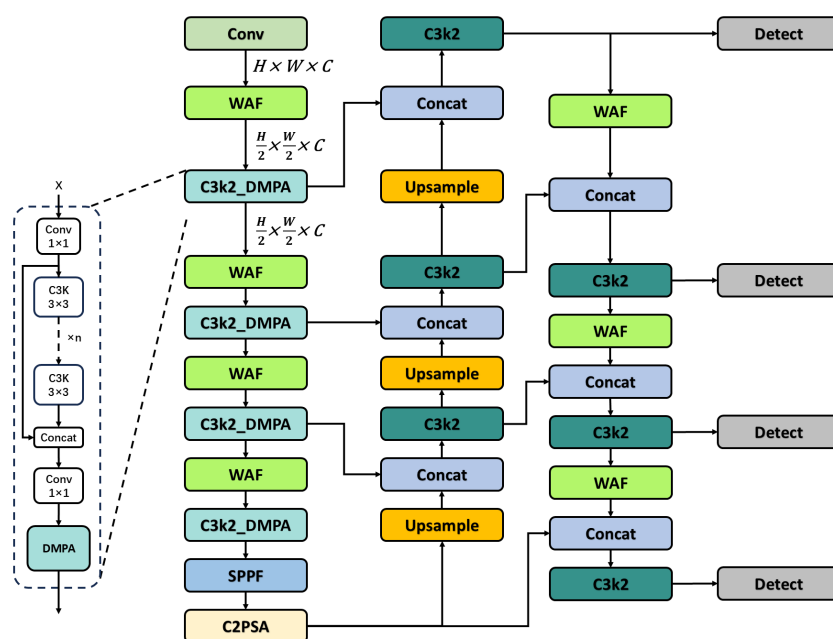


Figure 4. RWA-YOLO architecture. Based on YOLOv11, the C3k2 module is enhanced with a DMPA to strengthen feature extraction, while the original downsampling convolution is replaced by a WAF to better retain small-object details.

4.2.1. Design Motivation and Rationale

Indoor UAV detection in multi-view localization scenarios presents a set of challenges that are closely related to the frequency-domain characteristics of visual signals. Due to their compact physical size and relatively simple geometric structure, UAVs typically occupy only a small number of pixels in indoor images. As a result, their discriminative cues are primarily encoded in high-frequency components, such as edges, contours, and fine structural details. In contrast, variations in indoor illumination conditions mainly affect low-frequency image components, often leading to significant appearance changes while preserving the underlying object structure. This mismatch between target-relevant information and illumination-induced variations motivates the need for frequency-aware feature modeling in indoor UAV detection.

Conventional attention mechanisms widely used in object detection networks, such as Squeeze-and-Excitation (SE) blocks and Convolutional Block Attention Modules (CBAMs), focus on channel-wise or spatial-wise feature reweighting based on global pooling or convolutional operations. While effective for general object detection tasks, these mechanisms do not explicitly model frequency-domain information and may suppress localized

high-frequency responses that are critical for small-object perception. Transformer-based attention architectures provide powerful global context modeling capabilities; however, they typically introduce substantial computational overhead and rely on token aggregation or pooling operations, which may lead to over-smoothing effects and degraded sensitivity to small targets. These limitations make them less suitable for real-time indoor UAV localization, where both computational efficiency and fine-grained feature preservation are essential.

Motivated by the above observations, this work adopts a wavelet-aware design to explicitly decouple and model low- and high-frequency feature components. By integrating a discrete wavelet transform into the downsampling process, the proposed Wavelet-Aware Fusion (WAF) module reduces spatial resolution while preserving and selectively enhancing high-frequency details that correspond to small UAV structures. Compared with conventional pooling-based or strided convolution-based downsampling, this frequency-domain decomposition provides a more principled mechanism for retaining discriminative information under varying illumination conditions.

In addition, effective utilization of the preserved multi-frequency features requires a feature aggregation mechanism that can jointly capture local details and broader contextual cues. To this end, the proposed Dynamic Multi-scale Patch Attention (DMPA) module is designed to complement the wavelet-aware representation by enabling multi-scale receptive field expansion and patch-level feature aggregation. Unlike standard attention modules that operate on a single scale or rely on global statistics, DMPA dynamically integrates multi-scale information and enhances both local structural cues and global contextual awareness, which is particularly beneficial for small UAV detection in cluttered indoor environments.

Overall, the combination of wavelet-aware feature decomposition and dynamic multi-scale patch attention forms a tightly coupled architectural design that is explicitly aligned with the signal characteristics and practical constraints of indoor UAV detection. Rather than serving as a generic attention enhancement, the proposed RWA-YOLO architecture is specifically tailored to improve small-object perception and robustness to illumination variations while maintaining real-time performance requirements. Having clarified the overall design rationale, we next detail the specific architectural instantiations of the proposed DMPA and WAF modules.

4.2.2. Architecture of DMPA and WAF Modules

The proposed DMPA design is inspired by the PPA module in HCF-Net [26], with the goal of exploring more effective feature aggregation strategies for indoor UAV detection. The original PPA module enhances feature representation by combining convolution with spatial and channel attention. However, in the context of indoor UAV localization, the limited scale diversity and restricted receptive field of such designs may constrain their effectiveness when UAV targets appear at different scales and positions.

Motivated by these observations, we propose an improved Dynamic Multi-scale Patch Attention (DMPA) module, as illustrated in Figure 5. The DMPA introduces multi-scale dilated convolutions to expand the effective receptive field and employs a multi-scale, multi-prompt fusion structure termed Dynamic Patch Prompt Attention, which empirically enhances the modeling of both local details and global contextual information. In addition, the channel attention branch adopts the Hardsigmoid activation to improve sparsity and computational efficiency, and supports adaptive normalization between BN and GN to improve training robustness under the evaluated settings. Overall, the DMPA maintains a lightweight design while demonstrating improved detection performance in the considered indoor UAV localization scenarios.

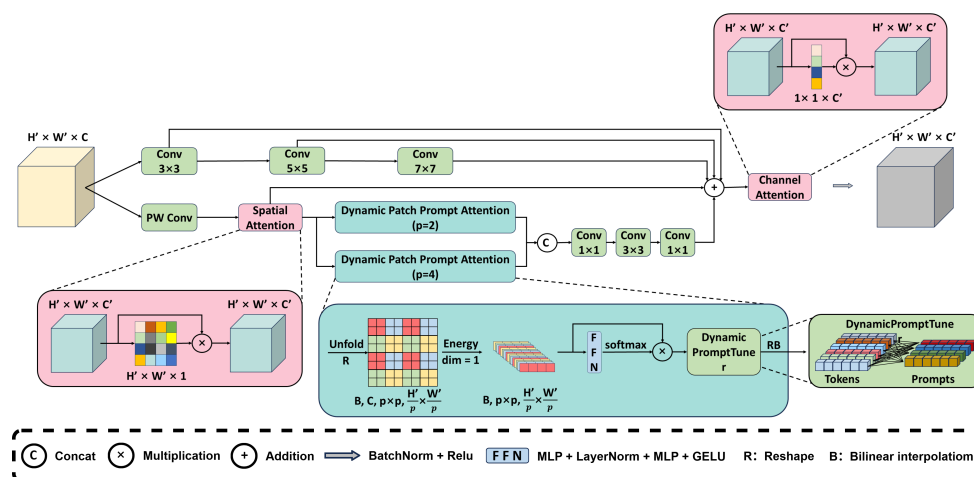


Figure 5. DMPA module. The proposed module fuses multi-scale features through convolutional blocks with different receptive fields. A Dynamic Patch Prompt Attention mechanism is incorporated to enhance both local and global feature representations. Moreover, spatial and channel attention modules are introduced to further strengthen the discriminative and expressive capability of the extracted features.

Since UAVs often occupy only a small number of pixels in an image, they are considered typical small targets. Consequently, the downsampling operations in the original YOLOv11 can result in the loss of small-object features, reducing recognition accuracy for distant UAVs. As illustrated in Figure 6, the proposed Wavelet-Aware Fusion (WAF) module integrates a discrete wavelet transform with an attention mechanism to alleviate this issue.

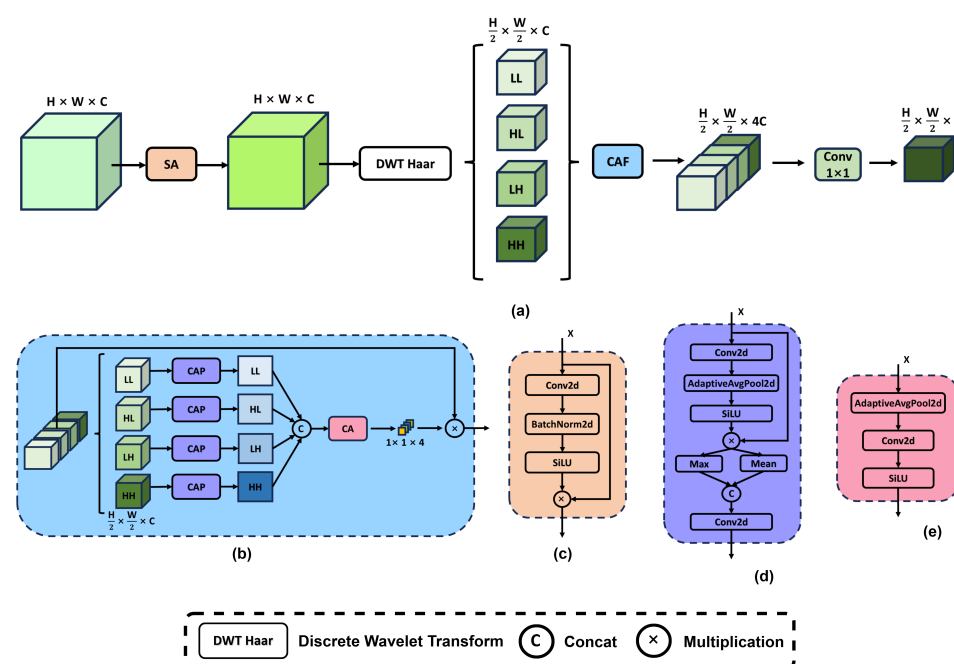


Figure 6. (a) Wavelet-Aware Fusion (WAF), (b) Channel Attention Fusion (CAF), (c) Spatial Attention (SA), (d) Channel Attention Pooling (CAP), and (e) Channel Attention (CA) modules. The Wavelet-Aware Fusion (WAF) module employs a one-level Haar-based discrete wavelet transform to perform feature decomposition and downsampling while preserving high-frequency information. By incorporating an attention mechanism, it enhances the discriminative representation of target features. Furthermore, the Channel Attention Fusion (CAF) module adaptively selects and integrates multi-frequency components derived from the wavelet transform, thereby maximizing the retention of salient target information.

Specifically, a one-level discrete wavelet transform (DWT) based on the Haar wavelet is applied to the input feature maps to decompose them into one low-frequency sub-band and three high-frequency sub-bands. The low-frequency component captures global structural information, while the high-frequency components preserve fine-grained details such as edges and contours of small UAV targets. Compared with conventional pooling-based downsampling, the wavelet-based decomposition enables spatial resolution reduction while retaining critical high-frequency information.

Subsequently, an attention mechanism is introduced to adaptively emphasize informative features and suppress irrelevant background responses. By jointly exploiting the multi-frequency representations provided by the Haar-based wavelet transform and the attention-guided feature refinement, the proposed WAF module effectively preserves discriminative details and enhances the detection performance of small UAV targets.

4.2.3. Robust UAV Localization Under Extreme Low-Light Conditions

Under favorable lighting conditions, the deep learning-based detection model can reliably output the UAV's 2D position. However, in low-light environments, image texture features degrade severely, and the recognition performance of the model declines significantly. To enhance detection robustness in low-light conditions, LEDs are mounted on the UAV body to provide localized illumination and increase image contrast. By introducing salient light intensity cues, the reliance on natural texture features is reduced, which facilitates more reliable detection in low-light indoor environments. The effectiveness of this design is experimentally validated in Section 5.

The lighting condition is evaluated on a per-frame basis using the mean grayscale image brightness. Specifically, the input image is first converted to grayscale, and the average pixel intensity is computed. When the mean brightness falls below a predefined threshold ($T_b = 40$ in our implementation), the environment is classified as low-light, and the system automatically switches from the deep learning-based detector to a brightness- and spatial-constraint-based light-spot detection method. Otherwise, the YOLO-based detector is applied. The threshold is empirically selected based on preliminary experiments, and the switching mechanism is fully automatic without requiring any manual presetting.

For UAV detection in low-light scenarios, a light-spot detection method based on brightness and spatial constraints is employed. First, a color-threshold segmentation is applied to extract high-intensity candidate regions, suppressing background noise. A minimum area constraint is then used to remove small noisy blobs. Since UAVs are small in size and their light spots occupy only limited pixels in the image, spatial clustering with an adaptive distance threshold is further applied to eliminate isolated light sources and reflection artifacts. The remaining candidate points are iteratively fused to obtain a stable set of UAV light spots, which effectively suppresses false detections and reduces frame-to-frame fluctuations.

The 2D UAV center point in each camera view is then calculated using:

$$(x_c, y_c) = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n y_i}{n} \right) \quad (8)$$

where (x_i, y_i) denote the 2D coordinates of each detected light spot, and n denotes the total number of detected light spots.

Finally, the computed 2D center point is combined with the camera projection matrix and integrated into the aforementioned multi-view geometry reconstruction method to estimate the UAV's 3D spatial position in low-light environments. During multi-view triangulation, geometrically inconsistent detections are further rejected based on reprojection error constraints, preventing false positives from propagating to the final localization

result. This design enables the proposed system to achieve continuous and robust UAV localization across all lighting conditions.

4.3. 3D Position Estimation

After obtaining the 2D pixel coordinates of the UAV in each camera view during the detection stage, the system converts these observations into 3D position information under a unified world coordinate system. To this end, a multi-view geometry model is adopted for spatial reconstruction. Since the intrinsic and extrinsic parameters of each camera have already been calibrated in the preprocessing stage, their corresponding projection matrices can be constructed. Based on these, a modified Direct Linear Transformation (DLT) algorithm is applied for 3D reconstruction.

DLT is essentially a linear solution method for estimating projection matrices. In this system, we use the known camera projection matrices $P_i \in \mathbb{R}^{3 \times 4}$ and the corresponding 2D image points $\mathbf{x}_i = (u_i, v_i, 1)^T$ to recover the UAV's 3D position in homogeneous world coordinates $\mathbf{X} = (X, Y, Z, 1)^T$. According to the projection model,

$$\mathbf{x} = P\mathbf{X} \quad (9)$$

which leads to the cross-product constraint

$$[\mathbf{x}_i]_{\times} (P_i \mathbf{X}) = 0 \quad (10)$$

Expanding (10) for each observation (u_i, v_i) gives two independent equations:

$$\begin{cases} u_i P_{i,3}^T \mathbf{X} - P_{i,1}^T \mathbf{X} = 0 \\ v_i P_{i,3}^T \mathbf{X} - P_{i,2}^T \mathbf{X} = 0 \end{cases} \quad (11)$$

which can be written compactly as

$$A\mathbf{X} = 0 \quad (12)$$

The matrix A is solved via Singular Value Decomposition (SVD):

$$A = U\Sigma V^T \quad (13)$$

and the solution $\mathbf{X}^*$ corresponds to the right singular vector associated with the smallest singular value:

$$\mathbf{X}^* = V_{:,4} \quad (14)$$

It should be noted that this SVD-based solution is applied to a calibrated multi-view triangulation problem with known camera parameters and metric scale. Under these conditions, the coefficient matrix A is well-conditioned, and the SVD yields a unique least-squares estimate of the UAV's 3D position. Therefore, the geometric ambiguities typically encountered in general Structure-from-Motion pipelines do not apply in this setup.

Finally, the Euclidean 3D coordinates of the UAV are obtained by normalization:

$$(X, Y, Z) = \frac{(X_1^*, X_2^*, X_3^*)}{X_4^*} \quad (15)$$

Since the system is equipped with four cameras, a minimum of two simultaneous observations of the UAV from different viewpoints is sufficient to recover its 3D position in the world coordinate system according to Equations (13)–(15). This approach ensures robust UAV localization even in the presence of partial occlusion or missing views.

Before 3D reconstruction, 2D detections with missing or low-confidence points are discarded. Only valid detections from at least two cameras are used for DLT-based triangulation. After initial triangulation, a weighted non-linear reprojection error optimization is applied to refine the 3D position. This process ensures that poor-quality views or partially occluded detections do not significantly degrade localization accuracy.

5. Experiments

5.1. Dataset Preparation

The object detection task aims to identify UAVs from multi-camera images and locate their two-dimensional bounding boxes, whose center coordinates are used as input for subsequent 3D position estimation. As shown in Figure 7, we collected a large number of UAV images in indoor environments from different viewpoints, distances, and scenes. These images were carefully annotated, with each sample labeled by a bounding box indicating the UAV position. To prevent the presence of humans in the scene from interfering with detection performance, we also annotated human instances in the dataset. In addition, to enhance the generalization capability of the model, we incorporated UAV images from public datasets to further enrich the training samples.

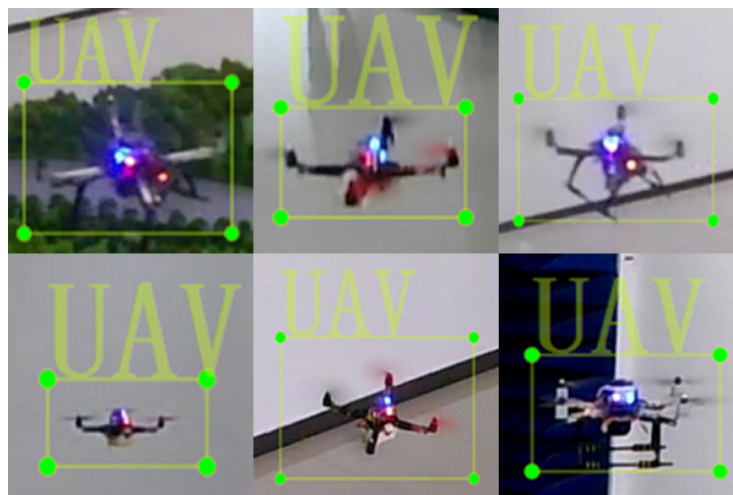


Figure 7. Examples of the UAV dataset.

For data processing, multiple data augmentation techniques were applied to improve model robustness. Specifically, image rotation, horizontal flipping, and vertical flipping were employed, which significantly increased the dataset size. The final dataset consists of approximately 12,000 annotated images. For model training and evaluation, the dataset was split into three subsets: 85% for training, 10% for validation, and 5% for testing. Considering the single-UAV scenario and the relatively consistent experimental environment in this study, the dataset scale and split ratio are sufficient to meet the expected detection accuracy requirements.

5.2. Experimental Settings

All experiments were conducted under identical environmental conditions as summarized in Table 1. Note that the RTX 4090 GPU and Core i9 CPU are only used during model training. Although all FPS results are reported on a desktop GPU, the lightweight architecture and moderate parameter size indicate strong potential for real-time deployment on embedded platforms, which will be validated in future work. Therefore, the cost of the deployed system remains low and practical for indoor UAV applications.

Table 1. Experimental Hardware and Software Configuration

Component	Specification
CPU	Intel Core i9-13900K
RAM	64 GB
GPU	NVIDIA GeForce RTX 4090 (24 GB VRAM)
CUDA Version	12.2
Operating System	Ubuntu Linux (x86_64)
Development Environment	Visual Studio Code

5.3. Evaluation Metrics

To comprehensively evaluate the performance of the proposed vision-based indoor positioning system for UAVs, several standard quantitative metrics were employed, including Precision (P), Recall (R), and mean Average Precision (mAP) under different Intersection over Union (IoU) thresholds. In addition, computational efficiency was assessed using the number of trainable parameters (Params) and the inference speed measured in frames per second (FPS).

To further evaluate the model's generalization capability under varying illumination conditions, the same metrics were computed on the UAV dataset. The definitions of the main evaluation metrics are as follows:

$$P = \frac{TP}{TP + FP} \quad (16)$$

$$R = \frac{TP}{TP + FN} \quad (17)$$

where TP (True Positives) denotes correctly detected objects, FP (False Positives) represents incorrect detections, and FN (False Negatives) indicates missed detections.

The Average Precision (AP) is defined as the area under the Precision–Recall curve:

$$AP = \int_0^1 P(R) dR \quad (18)$$

and the mean Average Precision (mAP) is obtained as:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (19)$$

where N represents the total number of object categories.

Specifically, $mAP@50$ denotes the mean AP at an IoU threshold of 0.5, while $mAP@50-95$ corresponds to the average over thresholds from 0.5 to 0.95 at 0.05 intervals. Furthermore, mAP_S , mAP_M , and mAP_L represent detection performance for small, medium, and large targets, respectively.

Model efficiency metrics are also reported, where $Params$ indicates the total number of trainable parameters, and FPS measures inference speed, serving as an indicator of real-time capability.

Finally, to assess robustness under illumination variations, $Precision_{UAV}$, $Recall_{UAV}$, $mAP@50_{UAV}$, and $mAP@50-95_{UAV}$ are computed using the UAV dataset.

5.4. Ablation Study

YOLOv11 was selected as the baseline due to its improved backbone efficiency and better compatibility with lightweight architectural modifications compared to earlier YOLO variants.

To evaluate the contribution of each proposed component, an ablation study was conducted using YOLOv11-Baseline as the reference model. The YOLOv11-Baseline corresponds to the official YOLOv11 architecture without any structural modifications. All baseline and ablated models are trained from scratch on the same indoor UAV dataset, without using external pre-trained weights, and are evaluated using identical training configurations and evaluation protocols.

In addition to module-level ablation, we further investigate the impact of different attention mechanisms to justify the architectural choice of the proposed DMPA. Specifically, the attention module embedded in the C3k2 block is replaced with representative alternatives, including Squeeze-and-Excitation (SE) and the Convolutional Block Attention Module (CBAM), while keeping the backbone, training strategy, and data preprocessing identical across all models.

As shown in Table 2, different attention mechanisms exhibit distinct performance characteristics under identical experimental settings. The SE module achieves a slightly higher $mAP@50$, indicating its effectiveness in enhancing coarse object detection through global channel reweighting. However, its improvement on small-object detection remains limited, as reflected by the modest gain in mAP_S .

Table 2. Comparison of different attention mechanisms under identical experimental settings. Bold values indicate the best value in each column.

Methods	P	R	$mAP@50$	$mAP@50-95$	mAP_S	Params	FPS
YOLOv11-Baseline	0.915	0.843	0.887	0.565	0.452	2.59M	27.47
YOLOv11 + SE	0.919	0.846	0.893	0.569	0.458	2.56M	27.61
YOLOv11 + CBAM	0.899	0.843	0.879	0.554	0.447	2.56M	32.69
YOLOv11 + DMPA (Ours)	0.952	0.810	0.889	0.575	0.473	5.84M	24.41

In contrast, the proposed DMPA achieves the highest $mAP@50-95$ and mAP_S among all compared methods, demonstrating superior performance in fine-grained localization and small UAV detection. Although DMPA introduces additional parameters and a moderate reduction in inference speed, its performance gains are disproportionately concentrated on metrics that are critical for indoor UAV detection, where targets occupy only a small number of pixels and require precise localization.

The inferior performance of the CBAM further suggests that generic spatial-channel attention may not effectively capture the discriminative patterns of small UAV targets in complex indoor environments. Overall, these results indicate that the performance improvements are not merely due to the presence of an attention mechanism, but are attributed to the proposed multi-scale patch-level aggregation strategy, which is better aligned with the requirements of small-object perception.

The overall detection performance under varying illumination conditions is reported in Table 3, while the efficiency and UAV-specific metrics are summarized in Table 4. The baseline model achieved an $mAP@50$ of 0.887 and an $mAP@50-95$ of 0.565, which serve as performance benchmarks.

Table 3. Overall detection performance of different modules for vision-based indoor UAV positioning under varying illumination. Bold values indicate the best value in each column.

Methods	P	R	$mAP@50$	$mAP@50-95$	mAP_S	mAP_M	mAP_L
YOLOv11-Baseline	0.915	0.843	0.887	0.565	0.452	0.621	0.622
+WAF	0.923	0.834	0.891	0.562	0.449	0.633	0.602
+DMPA	0.952	0.810	0.889	0.575	0.473	0.632	0.619
RWA-YOLO	0.925	0.836	0.898	0.586	0.472	0.646	0.639

Table 4. Efficiency and UAV-specific performance of different modules. Bold values indicate the best value in each column.

Methods	Params	FPS	P(UAV)	R(UAV)	mAP@50–95(UAV)
YOLOv11-Baseline	2.59M	27.47	0.934	0.945	0.621
+WAF	2.30M	29.24	0.948	0.955	0.633
+DMPA	5.84M	24.41	0.965	0.931	0.632
RWA-YOLO	5.55M	24.67	0.954	0.968	0.646

As shown in Table 3, incorporating the proposed *Wavelet-Aware Fusion* (WAF) block results in consistent improvements in precision (from 0.915 to 0.923) and $mAP@50$ (from 0.887 to 0.891), while simultaneously reducing the number of parameters from 2.59 M to 2.30 M. This indicates that the WAF module effectively enhances feature extraction efficiency without increasing model complexity. A slight decrease in $mAP@50–95$ is observed when WAF is applied alone. This does not indicate degraded detection capability; rather, it reflects the design emphasis of WAF on frequency-domain feature fusion and robustness enhancement, which improves detection stability and coarse localization, but may slightly smooth fine-grained spatial details required for high-IoU localization.

By introducing the Dynamic Multi-scale Patch Attention (DMPA) module, the $mAP@50–95$ improves from 0.565 to 0.575, reflecting enhanced multi-scale representation capability. However, as reported in Table 4, the additional attention mechanism increases model parameters to 5.84 M and slightly reduces inference speed from 27.47 FPS to 24.41 FPS, illustrating the trade-off between accuracy and computational efficiency.

When both modules are jointly integrated, the proposed RWA-YOLO effectively combines the complementary strengths of WAF and DMPA. Specifically, the patch-level spatial discrimination introduced by DMPA compensates for the slight loss of fine-grained localization caused by WAF, leading to the best overall performance.

As a result, RWA-YOLO achieves an $mAP@50$ of 0.898 and an $mAP@50–95$ of 0.586, corresponding to absolute improvements of 1.1 and 2.1 percentage points over the baseline, respectively. Moreover, as shown in Table 4, RWA-YOLO attains the highest UAV-specific precision (0.954), recall (0.968), and $mAP@50–95$ (0.646), demonstrating strong robustness and generalization capability under varying illumination conditions. In addition, consistent gains on small- and medium-scale objects further validate the effectiveness of the WAF and DMPA modules in enhancing feature representation and small-object perception.

5.5. Comparison with State-of-the-Art Methods

The proposed RWA-YOLO is compared with several state-of-the-art object detection models, including YOLOv5, YOLOv6, YOLOv8, YOLOX-L, DAMO-YOLO, RT-DETRv1-R101vd, and RT-DETRv2-R101vd. The comparison is conducted from two perspectives: overall detection performance and efficiency with UAV-specific accuracy. For fair comparison, all models were retrained on the same dataset using identical input resolution and training epochs. Default configurations recommended by the original authors were adopted whenever applicable.

Table 5 reports the overall detection results in terms of precision (P), recall (R), $mAP@50$, $mAP@50–95$, and scale-specific mAP values. Among all compared methods, RWA-YOLO achieves the highest $mAP@50$ (0.898) and $mAP@50–95$ (0.586), demonstrating superior overall detection accuracy. Notably, it also attains the best performance on small objects with an mAP_s of 0.472, which is particularly important for indoor UAV localization scenarios where the UAV often occupies limited image regions.

Table 6 further presents the model complexity, inference speed, and UAV-specific evaluation results. Although RWA-YOLO introduces a moderate increase in parameters

(5.55 M) compared with lightweight YOLO variants, it remains significantly more compact than transformer-based RT-DETR models while maintaining near real-time inference speed (24.67 FPS for neural network inference). The subsequent multi-view triangulation step, which computes the UAV's 3D position from multiple camera views, introduces negligible computational overhead (<1 ms per frame). Therefore, the total processing time per frame, including both neural network inference and triangulation, corresponds to an effective FPS of approximately 24.7, demonstrating that the system maintains real-time performance for indoor UAV localization tasks.

Table 5. Overall detection performance comparison of different object detection models. Bold values indicate the best value in each column.

Model	P	R	mAP@50	mAP@50–95	mAP _S	mAP _M	mAP _L
YOLOv5	0.897	0.848	0.883	0.548	0.441	0.588	0.616
YOLOv6	0.902	0.832	0.877	0.557	0.455	0.607	0.610
YOLOv8	0.925	0.848	0.895	0.566	0.461	0.621	0.616
YOLOX-L	0.880	0.592	0.883	0.514	0.186	0.486	0.569
DAMO-YOLO	0.512	0.626	0.860	0.512	0.159	0.466	0.576
RT-DETRv1-R101vd	0.934	0.758	0.934	0.637	0.290	0.587	0.710
RT-DETRv2-R101vd	0.932	0.765	0.932	0.637	0.275	0.585	0.712
RWA-YOLO (Ours)	0.925	0.836	0.898	0.586	0.472	0.646	0.639

Table 6. Efficiency and UAV-specific performance comparison of different object detection models. Bold values indicate the best value in each column.

Model	Params	FPS	Precision(UAV)	Recall(UAV)	mAP@50–95(UAV)
YOLOv5	2.51M	28.07	0.929	0.941	0.588
YOLOv6	4.23M	33.15	0.946	0.927	0.607
YOLOv8	3.01M	34.36	0.953	0.941	0.621
YOLOX-L	54M	124	0.450	0.536	0.453
DAMO-YOLO	27.56M	122	0.415	0.528	0.415
RT-DETRv1-R101vd	76.37M	79.54	0.863	0.616	0.559
RT-DETRv2-R101vd	76.37M	85.65	0.863	0.630	0.551
RWA-YOLO (Ours)	5.55M	24.67	0.954	0.968	0.646

In terms of UAV-specific performance, RWA-YOLO consistently outperforms all competing methods, achieving the highest precision (0.954), recall (0.968), and *mAP*@50–95 (0.646). These results validate the effectiveness of the proposed wavelet-aware feature fusion and dual-mode parallel attention mechanisms in enhancing multi-scale representation and small-object perception.

Furthermore, we evaluate our vision-based framework against representative hybrid localization methods reported in the literature to contextualize our accuracy. While hybrid systems like VIUNet [14] utilize ultra-wideband (UWB) sensors to assist visual detection, they often require complex infrastructure and periodic calibration. Our RWA-YOLO achieves a centimeter-level localization RMSE of 9.9 mm, which is competitive with, or even superior to, the reported precision of such hybrid UWB-vision frameworks. This demonstrates that by leveraging wavelet-aware feature fusion, our pure vision-based approach can achieve high-fidelity localization without the hardware overhead and synchronization challenges associated with multi-modal sensor fusion.

Overall, RWA-YOLO strikes a favorable balance between accuracy, efficiency, and inference speed, making it well suited for real-time indoor UAV localization tasks.

5.6. Discussion on Comparison with Visual SLAM and VO

Visual SLAM and visual odometry are widely adopted solutions for indoor UAV localization. However, a direct quantitative comparison with the proposed system is inherently constrained by fundamental differences in system design, sensing configuration, and evaluation protocols. Representative SLAM/VO methods, such as ORB-SLAM3 and VINS-Mono, estimate relative poses using onboard sensors and are commonly evaluated in terms of trajectory-level accuracy over time, e.g., absolute trajectory error (ATE) or relative pose error (RPE), under unified experimental settings.

In contrast, the proposed system estimates the UAV's absolute 3D position at discrete time instances based on external multi-view observations, and is evaluated using point-wise localization error with respect to predefined control points in a global coordinate frame. As a result, the output representation, coordinate reference, and performance metrics differ substantially between the two paradigms, making a direct metric-level comparison under an identical protocol inappropriate.

Nevertheless, situating the proposed approach with respect to SLAM/VO systems remains important, and a qualitative comparison helps clarify their complementary roles. SLAM and VO methods rely on incremental motion estimation and are therefore susceptible to accumulated drift over long trajectories or in texture-poor indoor environments, even with loop closure mechanisms. In contrast, the proposed method provides drift-free absolute positioning, as each position estimate is independently computed from external visual constraints.

From a system-level perspective, the proposed approach can be regarded as an absolute positioning counterpart to SLAM-based relative localization. Rather than replacing SLAM or VO methods, it complements them by providing global position references for system initialization or periodic drift correction. This characteristic is particularly advantageous for indoor UAV applications that require long-term global consistency or operate under strict onboard computational and power constraints.

5.7. System Evaluation

All localization experiments are conducted with respect to a predefined and fixed world coordinate system. The world coordinate origin is manually defined within the experimental environment and remains unchanged throughout all measurements. The three-dimensional coordinates of all control points are expressed in this world coordinate frame and serve as ground truth references for localization accuracy evaluation.

All evaluations in this section are conducted under static conditions, where the UAV is placed at fixed locations corresponding to the predefined control points, in order to isolate and assess the intrinsic localization accuracy and stability of the proposed system.

To comprehensively evaluate the proposed vision-based indoor UAV localization system, a set of spatial control points with known world coordinates is selected as test samples. For each control point, the UAV is positioned at the corresponding location, and its estimated 3D position is compared against the ground-truth coordinate expressed in the same world coordinate system. To compensate for potential global translation, rotation, and scale discrepancies between the estimated coordinate frame and the reference frame, the predicted UAV positions are aligned to the world coordinate system using the Umeyama algorithm. This alignment process does not alter the relative localization errors and ensures a fair and consistent quantitative evaluation.

The overall experimental performance of the proposed system is illustrated in Figure 8. Localization accuracy is quantified using multiple complementary metrics, including the root mean square error (RMSE) of the 3D position, the 95th percentile error (P95), component-wise RMSE on the XY, XZ, and YZ planes, the leave-one-out mean error

(LOO_Mean), the bootstrap-based 95% confidence interval (CI) of the localization error, and the estimated scale factor between the predicted and ground-truth coordinates. These metrics collectively provide a comprehensive assessment of both localization accuracy and system stability.

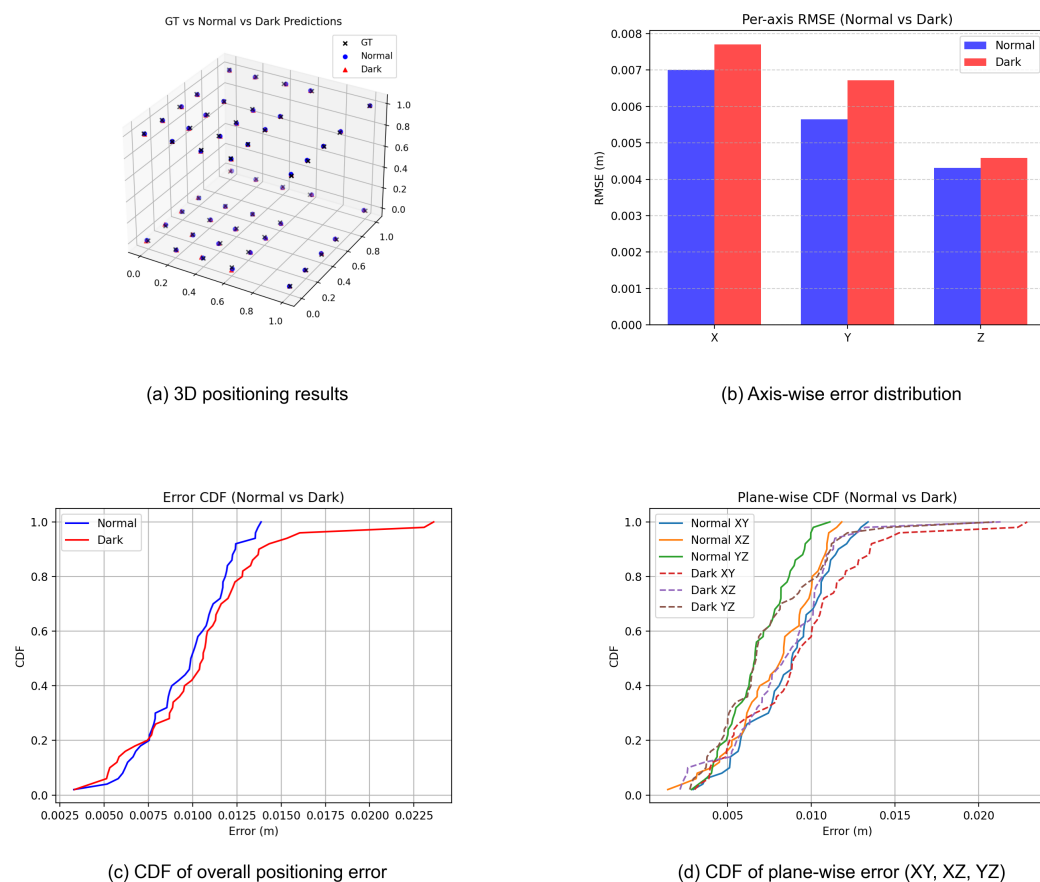


Figure 8. Overall experimental results of the proposed indoor UAV localization system.

Table 7 summarizes the quantitative localization results under normal and low-light illumination conditions. Under normal lighting, the proposed system achieves centimeter-level localization accuracy, with an RMSE of approximately 0.01 m and a P95 error of approximately 0.013 m. The low component-wise RMSE values indicate well-balanced accuracy across different spatial directions. Even in low-light environments, the localization accuracy only exhibits a slight degradation, with an RMSE of approximately 0.011 m and a P95 error of approximately 0.016 m, which remains sufficient for stable indoor UAV localization.

Table 7. Quantitative evaluation of UAV localization accuracy under normal and low-light illumination conditions.

Metric	Normal (m)	Low-Light (m)
RMSE	0.0100	0.0112
P95	0.0135	0.0157
RMSE XY/XZ/YZ	0.0090, 0.0082, 0.0071	0.0102, 0.0090, 0.0081
LOO_Mean	0.0102	0.0110
Bootstrap 95% CI	[0.0100, 0.0105]	[0.0113, 0.0118]
Scale	1.0035	1.0007

As shown in Figure 9, the proposed vision-based detector reliably identifies the UAV across all illumination conditions. In particular, under low-light environments, the onboard

LED light sources generate visually salient high-intensity regions that remain clearly distinguishable from the background, thereby facilitating reliable UAV detection when natural texture features are severely degraded. These results confirm the robustness of the perception module under challenging lighting variations. Overall, the results demonstrate that the proposed method provides both high-precision localization and strong robustness without reliance on costly external hardware. The system achieves real-time performance with controllable computational complexity, highlighting its practical feasibility for indoor UAV localization applications.

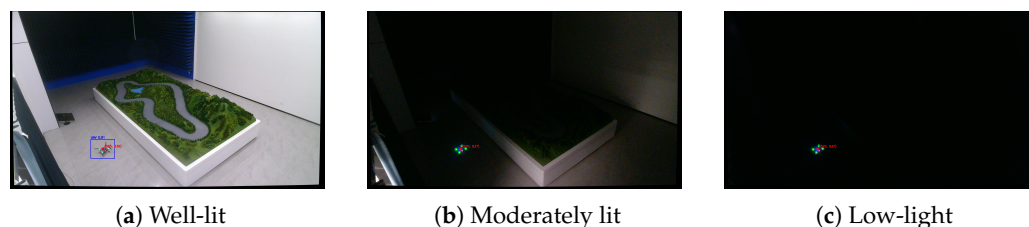


Figure 9. UAV detection results under different illumination conditions: (a) well-lit, (b) moderately lit, and (c) low-light environments. The proposed vision-based method maintains reliable UAV detection across all scenarios, demonstrating strong robustness to illumination variations.

5.8. Dynamic Trajectory Evaluation and Path Reconstruction

To further validate the practical utility of RWA-YOLO for real-time UAV navigation, we conducted a dynamic flight experiment to assess the system's capability for continuous path reconstruction. Unlike static evaluations at discrete control points, this experiment requires the detector to maintain high precision and temporal consistency under motion-induced conditions.

5.8.1. Experimental Setup for Dynamic Flight

A quadrotor UAV was manually piloted to follow a predefined flight pattern (approximately rectangular, while maintaining a certain flight altitude) within the $10\text{ m} \times 8\text{ m} \times 5\text{ m}$ workspace. For each frame, the RWA-YOLO detector identified the UAV's 2D position across available views, and the 3D coordinates were reconstructed via the triangulation pipeline described in Section 4.

5.8.2. Trajectory Consistency and Smoothness

As shown in Figure 10, the reconstructed trajectory exhibits smooth and continuous motion throughout the flight. While no ground-truth trajectory is available for direct comparison, the position estimates remain consistent across frames, without noticeable jumps or outliers.

To quantify temporal stability, we measured the frame-to-frame displacement variance and found it remained within a centimeter-level range. This indicates that the system provides reliable and temporally coherent position estimates.

5.8.3. Real-Time Performance for Control Loops

The effective update rate of the entire localization pipeline (inference + triangulation) reached approximately 24.7 FPS, corresponding to a latency of roughly 40.5 ms. Considering that most high-level position control loops operate at 10–20 Hz, this demonstrates that the proposed system can provide sufficiently frequent and smooth position feedback for stable indoor UAV operation.

It should be noted that this experiment is intended to validate the localization module under dynamic motion. The design and evaluation of closed-loop autonomous flight control are beyond the scope of this work.

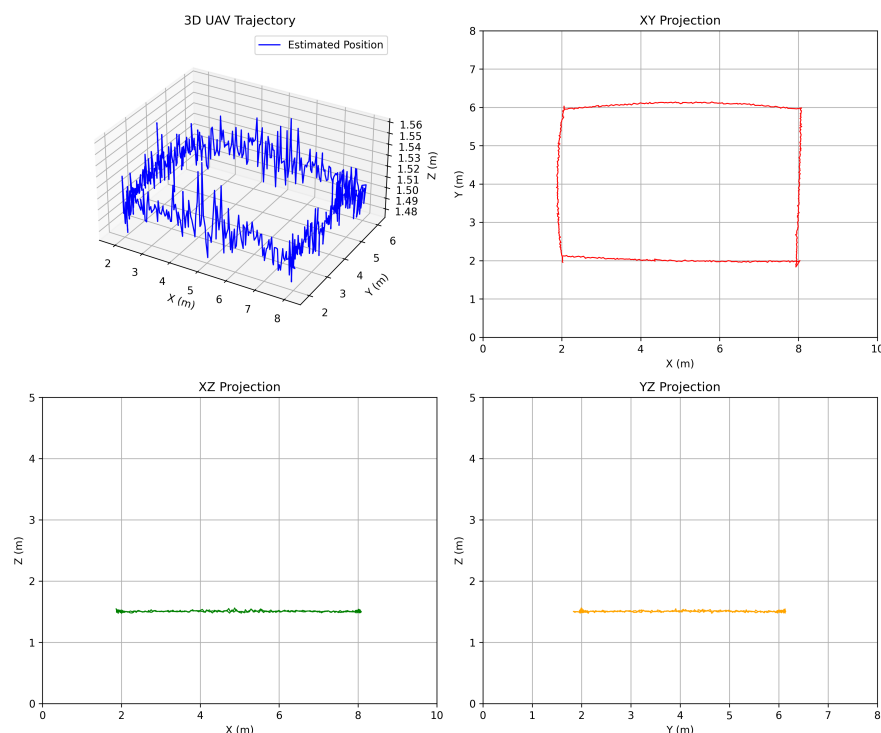


Figure 10. Reconstructed UAV trajectory during manual flight. The trajectory demonstrates spatial continuity and smoothness. Projections on the XY, XZ, and YZ planes indicate consistent position estimates across different views.

5.9. Evaluation Under Practical Deployment Conditions

While the above experiments demonstrate high localization accuracy under a well-controlled multi-camera setup, real-world indoor deployments may involve additional challenges such as sparse camera placement, partial visual occlusions, and extended spatial coverage. To further assess the generalizability and practical applicability of the proposed system, additional experiments were conducted under these more challenging conditions.

5.9.1. Impact of Camera Density and Redundancy

In this work, camera density refers to the total number of cameras deployed in the environment, while camera redundancy describes whether the UAV is simultaneously observed by more than the minimum two camera views required for 3D triangulation. Although two views are sufficient for reconstructing a 3D position, additional views introduce redundancy, providing extra geometric constraints that improve robustness against occlusion, temporary detection errors, and unfavorable viewing angles.

To investigate how the proposed system behaves as such redundancy is reduced, we systematically vary the number of active camera views used during the triangulation stage. In practical indoor deployments, the number of available cameras may be limited by installation cost, physical constraints, or environmental layout. To emulate these conditions without modifying the physical setup, camera sparsity is simulated by selectively masking camera streams during reconstruction, while keeping the detection network, calibration parameters, and reconstruction pipeline unchanged. Localization performance is evaluated using four, three, and two active cameras.

Table 8 summarizes the localization accuracy under different camera configurations. As expected, localization precision gradually degrades as the number of active views decreases, due to reduced geometric constraints and increased sensitivity to 2D detection noise. With four cameras, the system achieves the best performance (lowest RMSE and P95 error), benefiting from maximum multi-view redundancy. When reduced to three cameras,

the localization error increases only marginally and remains within the centimeter-level range, indicating strong tolerance to partial view loss.

Even under the minimal two-camera configuration required for triangulation, the system continues to provide stable 3D position estimates, albeit with a predictable increase in uncertainty. Geometric consistency is enforced through reprojection error checks, and unreliable reconstructions are discarded. Overall, these results demonstrate that the proposed framework exhibits graceful performance degradation rather than catastrophic failure under sparse camera deployments, highlighting its robustness and suitability for real-world indoor environments where dense camera coverage may not be feasible.

Table 8. Localization performance under varying camera availability (simulated occlusion).

Configuration	Active Views	RMSE (m)	P95 (m)
Four Cameras (Baseline)	4	0.0100	0.0135
Three Cameras	3	0.0124	0.0172
Two Cameras	2	0.0187	0.0264
Single View Case	1	N/A	N/A

When only two cameras are visible, the system operates at the minimum configuration required for 3D reconstruction. Geometric consistency is enforced by evaluating reprojection errors, and unreliable reconstructions are discarded. As reported in Table 8, although the positioning accuracy slightly decreases under the two-camera configuration, the system remains stable and does not exhibit significant error amplification.

5.9.2. Robustness to Partial Occlusions

In realistic indoor environments, the UAV may be partially occluded by furniture, shelves, or structural elements, resulting in unreliable or missing observations in certain camera views. Unlike the camera density study in the previous subsection, which considers permanent reductions in available viewpoints, partial occlusions represent transient and view-dependent observation failures.

To evaluate robustness under such conditions, partial occlusion scenarios were simulated at the observation level. Specifically, during 3D reconstruction, individual 2D detections with low confidence or large reprojection residuals were selectively discarded, emulating situations where the UAV is temporarily occluded or poorly observed in specific camera views. The remaining valid views were then used for triangulation without modifying the detection network, calibration parameters, or reconstruction pipeline.

Experimental results indicate that the proposed system maintains accurate and stable localization as long as sufficient valid observations are available. Although partial occlusions introduce additional uncertainty due to reduced effective viewpoints, the resulting increase in localization error remains bounded. This behavior demonstrates that the multi-view fusion strategy can tolerate intermittent observation loss and effectively suppress the influence of unreliable views.

Overall, the system exhibits graceful degradation rather than catastrophic failure under partial occlusions, which is critical for deployment in cluttered indoor environments where complete visibility from all cameras cannot be guaranteed.

5.9.3. Extended Spatial Coverage

The experiments were conducted in a real indoor environment with approximate dimensions of 10 m × 8 m × 5 m (length × width × height), which provides a non-trivial spatial scale for evaluating the proposed system. This setup introduces multi-meter camera-to-UAV distances, depth variations, and viewpoint changes that are representative of

practical indoor deployment scenarios such as laboratories, industrial rooms, and medium-sized halls.

As the workspace expands, the average distance between the UAV and the camera network increases, leading to reduced target resolution in the image plane and increased sensitivity of 3D triangulation to pixel-level localization errors. Despite these challenges, the proposed system consistently achieves centimeter-level localization accuracy across the evaluated workspace, as demonstrated by the quantitative results reported in Table 7.

From a geometric perspective, the observed error characteristics align with multi-view projection theory: localization uncertainty grows gradually with distance due to diminishing parallax and amplified reprojection noise. Notably, no abrupt performance degradation or instability was observed within the tested spatial range, indicating that the proposed system maintains stable operation under realistic indoor scales.

While the current evaluation focuses on a medium-sized indoor environment, the proposed framework is not inherently limited to this scale. Extension to larger indoor spaces can be achieved by increasing camera baselines, deploying additional views, or partitioning the workspace into overlapping sub-regions. These strategies preserve sufficient geometric constraints and ensure robust triangulation performance. Investigation of optimal camera layouts for large-scale indoor environments is left as future work.

5.10. Sources of Localization Error

The localization accuracy of the proposed multi-view vision-based system is influenced by multiple error sources along the measurement chain, from 2D image observation to 3D geometric reconstruction. This subsection analyzes the primary sources of localization error and discusses their impact on the final 3D positioning accuracy.

The main sources of localization error can be summarized as follows:

- Pixel-level 2D localization error: inaccuracies in estimating the UAV center position in each image, caused by detection errors, image noise, and illumination variations;
- Camera calibration error: residual inaccuracies in intrinsic and extrinsic parameters obtained from camera calibration;
- Geometric configuration: the spatial arrangement of cameras, including baseline length and viewing angles, which affects triangulation sensitivity.

Among these factors, the pixel-level 2D localization error is considered the dominant source in the proposed system, as it directly propagates to the 3D reconstruction through the multi-view triangulation process.

To quantitatively evaluate the influence of pixel-level localization error, Gaussian noise with different standard deviations was artificially added to the detected 2D UAV center coordinates before triangulation. Specifically, noise levels of 0.5 px, 1.0 px, and 2.0 px were considered to simulate typical detection uncertainties under different imaging conditions. The noisy 2D observations were then used for 3D position estimation using the same triangulation pipeline.

Table 9 reports the resulting 3D localization root mean square error (RMSE) under different pixel-level noise levels. As can be observed, the 3D localization error increases almost linearly with respect to the pixel-level measurement uncertainty, indicating a strong sensitivity of the triangulation accuracy to 2D detection precision.

Table 9. Effect of pixel-level localization error on 3D positioning accuracy.

Pixel Noise (px)	3D RMSE (mm)
0.5	6.0
1.0	10.0
2.0	19.5

Although a complete analytical uncertainty propagation model is beyond the scope of this work, the above analysis and experimental results demonstrate that the proposed system achieves a stable centimeter-level localization uncertainty under the tested conditions. The improved robustness of the proposed detector under varying illumination effectively reduces pixel-level measurement uncertainty, thereby enhancing the overall reliability of the multi-view localization system. Future work will focus on establishing a comprehensive uncertainty model for the proposed measurement framework.

6. Conclusions

We present RWA-YOLO, a vision-based indoor UAV localization framework achieving centimeter-level accuracy without external beacons or fiducial markers. The system leverages a multi-camera setup and an enhanced YOLO detector with wavelet-aware attention and dual multi-path aggregation modules, improving small-object detection and feature representation under challenging lighting and occlusion conditions. Extensive experiments in real indoor environments demonstrate a 3D positioning RMSE of 9.9 mm and P95 accuracy of 13.5 mm, demonstrating performance that outperforms YOLOv11 and is comparable to or better than representative hybrid localization methods reported in the literature.

While the current work focuses on high-precision 3-DoF absolute localization (x, y, z) , which is critical for suppressing the long-term drift in GPS-denied environments, future work will extend the framework to 6-DoF pose estimation. By identifying the geometric configurations of multiple onboard markers, the system could provide complete orientation information. Additionally, we aim to further optimize the computational efficiency for deployment on resource-constrained embedded platforms and explore deep integration with onboard IMU data to enhance robustness during high-speed maneuvers.

These results confirm that RWA-YOLO provides robust and accurate indoor UAV localization under varying illumination and limited texture, demonstrating its practical applicability in real-world scenarios. It is worth noting that the proposed external vision-based localization framework is complementary to onboard SLAM/VO systems and can potentially be integrated as an absolute position reference to correct long-term drift in autonomous UAV navigation.

Author Contributions: Conceptualization, F.W., Y.W.; Methodology, F.W., K.S.; Software, F.W.; Validation, F.W.; Formal analysis, F.W.; Investigation, F.W., K.S.; Resources, Y.W.; Data curation, F.W. and K.S.; Writing—original draft preparation, F.W.; Writing—review and editing, F.W.; Visualization, F.W.; Supervision, Y.W.; Project administration, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study was collected in a controlled indoor environment using a specific multi-camera setup. Due to confidentiality and institutional restrictions, the full dataset is not publicly available. However, representative data samples and detailed descriptions of the data acquisition and experimental setup are provided in the manuscript. The dataset can be made available from the corresponding author upon reasonable request for research purposes.

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

References

- Lin, H.Y.; Zhan, J.R. GNSS-Denied UAV Indoor Navigation with UWB Incorporated Visual Inertial Odometry. *Measurement* **2023**, *212*, 112256. [\[CrossRef\]](#)
- Yang, B.; Yang, E.; Yu, L.; Loeliger, A. High-Precision UWB Based Localisation for UAV in Extremely Confined Environments. *IEEE Sens. J.* **2022**, *22*, 14435–14444. [\[CrossRef\]](#)
- Tiemann, J.; Wietfeld, C. Scalable and Precise Multi-UAV Indoor Navigation Using TDOA-Based UWB Localization. In Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sapporo, Japan, 18–21 September 2017.
- Ledergerber, A.; Hamer, M.; D’Andrea, R. A Robot Self-Localization System Using One-Way Ultra-Wideband Communication. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 3131–3137.
- Marins, J.; Yun, X.; Bachmann, E.; McGhee, R.; Zyda, M. An Extended Kalman Filter for Quaternion-Based Orientation Estimation Using MARG Sensors. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Maui, HI, USA, 29 October–3 November 2001; pp. 2003–2011.
- Faragher, R.; Harle, R. An Analysis of the Accuracy of Bluetooth Low Energy for Indoor Positioning Applications. In Proceedings of the ION GNSS+, Tampa, FL, USA, 8–12 September 2014; pp. 201–210.
- He, S.; Chan, S. Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 466–490. [\[CrossRef\]](#)
- Garrido-Jurado, S.; Muñoz-Salinas, R.; Madrid-Cuevas, F.; Marín-Jiménez, M. Automatic Generation and Detection of Highly Reliable Fiducial Markers under Occlusion. *Pattern Recognit.* **2014**, *47*, 2280–2292. [\[CrossRef\]](#)
- Mur-Artal, R.; Montiel, J.; Tardós, J. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [\[CrossRef\]](#)
- Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-Based Visual-Inertial Odometry Using Nonlinear Optimization. *Int. J. Robot. Res.* **2015**, *34*, 314–334. [\[CrossRef\]](#)
- Habaebe, M.; Khamis, R.; Islam, M. Mobile Drone Localization in Indoor Environment Based on Passive RFID. *Int. J. Interact. Mob. Technol.* **2020**, *14*, 4–15. [\[CrossRef\]](#)
- Loquercio, A.; Kaufmann, E.; Ranftl, R.; Dosovitskiy, A.; Koltun, V.; Scaramuzza, D. Deep Drone Racing: Learning Agile Flight in Dynamic Environments. *IEEE Trans. Robot.* **2020**, *36*, 1–20. [\[CrossRef\]](#)
- Schmid, K.; Tomic, T.; Ruess, F.; Hirschmüller, H.; Suppa, M. Stereo Vision-Based Indoor/Outdoor Navigation for Flying Robots. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 215–222.
- Kao, P.Y.; Chang, H.J.; Tseng, K.W.; Chen, T.; Luo, H.L.; Hung, Y.P. VIUNet: Deep Visual-Inertial-UWB Fusion for Indoor UAV Localization. *IEEE Access* **2023**, *11*, 34567–34578. [\[CrossRef\]](#)
- Xiang, Z.; Chen, L.; Wu, Q.; Yang, J.; Dai, X.; Xie, X. An Improved UWB Indoor Positioning Approach for UAVs Based on the Dual-Anchor Model. *Sensors* **2025**, *25*, 1052. [\[CrossRef\]](#) [\[PubMed\]](#)
- Bu, S.; Zhang, J.; Li, X.; Li, K.; Hu, B. IVU-AutoNav: Integrated Visual and UWB Framework for Autonomous Navigation. *Drones* **2025**, *9*, 162. [\[CrossRef\]](#)
- Qin, T.; Li, P.; Shen, S. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [\[CrossRef\]](#)
- Hsieh, T.L.; Jhan, Z.S.; Yeh, N.J.; Chen, C.Y.; Chuang, C.T. An Unmanned Aerial Vehicle Indoor Low-Computation Navigation Method Based on Vision and Deep Learning. *Sensors* **2024**, *24*, 190. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hu, Z.; Butt, M.; Nasir, N.; Salleh, M. Vision-Based Autonomous Navigation for Quadrotor UAVs in Unknown and Constrained Environments. *Int. J. Intell. Robot. Appl.* **2025**, *9*, 1743–1754. [\[CrossRef\]](#)
- Yang, J.; Tian, J.; Qi, Y.; Cheng, W.; Liu, Y.; Han, G.; Wang, S.; Li, Y.; Cao, C.; Qin, S. Research on 3D Localization of Indoor UAV Based on Wasserstein GAN and Pseudo Fingerprint Map. *Drones* **2024**, *8*, 740. [\[CrossRef\]](#)
- Krupáš, M.; Pačuta, K.; Zolotová, I. Towards Autonomous UWB-Based UAV Indoor Navigation. *Acta Electrotech. Inform.* **2025**, *25*, 10–14. [\[CrossRef\]](#)
- Yin, C.; Wang, C.; Chen, J.; Jiang, H.; Miao, X.; Zheng, S.; Yan, H. CiUAV: A Multi-Task 3D Indoor Localization System for UAVs Based on Channel State Information. *arXiv* **2025**, arXiv:2505.21216.
- Choutri, M.; Kerrar, A.; Guessoum, A.; Hamdi, F.; Alimi, A. Vision-Based UAV Detection and Localization to Indoor Positioning System. *Sensors* **2024**, *24*, 4121. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shalaby, M.; Ahmed, S.; Dahdah, N.; Cossette, C.; Le Ny, J.; Forbes, J. MILUV: A Multi-UAV Indoor Localization Dataset with UWB and Vision Measurements. *arXiv* **2025**, arXiv:2504.14376. [\[CrossRef\]](#)

25. Zimmerman, N.; Müller, H.; Magno, M.; Benini, L. Fully Onboard Low-Power Localization with Semantic Sensor Fusion on a Nano-UAV Using Floor Plans. *arXiv* **2023**, arXiv:2310.12536.
26. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Teng, X.; Li, A.; Guo, L. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. In Proceedings of the IEEE International Conference on Multimedia Expo (ICME), Bangkok, Thailand, 6–10 July 2024; pp. 1–6.

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.