Learning Object Following for Small UAVs using Vision-Language and Efficient Visual Flight Control

Anonymous Author(s)
Affiliation
Address
email

Abstract

Indoor drone navigation is challenging due to GPS denial and sensor drift. We introduce a vision-language-guided framework where a frozen model generates pseudo-labels to train a lightweight YOLOv8 student for markerless drone detection. Coupled with a depth-free IBVS controller operating solely on 2D image features, our system enables robust, real-time indoor UAV tracking with minimal manual supervision. Validated in classroom flights, it performs reliably across varied trajectories and supports scalable deployment for surveillance, inspection, and automation. A demonstration is available at: `https://youtu.be/67IEGqvVRYI`.

1 Introduction

Autonomous UAVs are increasingly used in applications such as industrial inspection [26, 9] and aerial cinematography [14], relying on object detection and tracking for navigation. While conventional systems using IMUs, GPS, and 9-DOF sensors perform well outdoors [1], they struggle in indoor and GPS-denied environments due to signal loss and sensor drift [24, 17].Vision-based navigation provides an alternative by extracting real-time features using onboard cameras [7, 6]. However, most drone-to-drone detection studies focus on outdoor settings, leaving indoor environments underexplored despite their need for precise relative tracking. Traditional Image-Based Visual Servoing (IBVS) methods often rely on depth estimation and suffer from numerical instabilities under large camera motion. To address this, we propose a vision-language-based UAV tracking framework combining self-supervised object detection with a depth-free IBVS controller.

Our main contributions are:

- **Vision-language teacher-student framework**: We train a YOLOv8 student detector using pseudo-labels generated by a frozen vision-language model (GroundingSAM), enabling scalable, self-supervised annotation for drone detection.
- **Depth-free visual servoing controller**: We design a reduced 4-DoF IBVS controller that operates exclusively on 2D image features, avoiding the instability and hardware requirements of depth estimation.
- **Curated indoor dataset with high-fidelity pseudo-labels**: We contribute a high-resolution indoor aerial dataset annotated automatically with over 98% accuracy, preserving temporal structure from video and supporting domain-specific detection tasks.

2 Related Work

Recent aerial detection efforts primarily target outdoor scenarios. Sangam et al. [20] use transformer-based spatio-temporal models, which are computationally heavy for UAVs. Reis et al. [18] propose lightweight YOLOv8 variants, but their datasets lack indoor and drone-to-drone examples. Ashraf et al. [2] similarly focus on outdoor dogfighting scenarios. For target following, visual servoing enables feedback-based control from visual inputs. Position-Based Visual Servoing (PBVS) recovers full 3D pose [16, 23], while Image-Based Visual Servoing (IBVS) operates directly on image features [4, 27, 11]. IBVS is preferred for real-time use due to its lower computational cost [28, 10]. Marker-based tracking methods usin ArUco [8, 13, 5] and AprilTags [15, 29, 3] are common but sensitive to lighting, occlusion,

Submitted to 2nd Best Romanian AI Thesis Awards (BRAIT 2025). Do not distribute.

and require setup. Prior work includes moving target tracking [12] and PID-based control [25], but these solutions are less viable in dynamic or unstructured environments. This motivates our markerless, vision-language-guided approach for indoor UAV tracking.

3 Approach

3.1 Dataset

Addressing the outdoor bias identified by Reis et al. [18] and Ashraf et al. [2], we collected indoor aerial data using two drones operated by human pilots in a controlled classroom environment. The drones flew in circular patterns around the classroom containing various objects, generating 4 videos (2 collected by the drone, 2 collected with a smartphone around the drones) from which we sampled 1 frame per second, totaling 14,400 frames with median image ratio of 3840×2160.

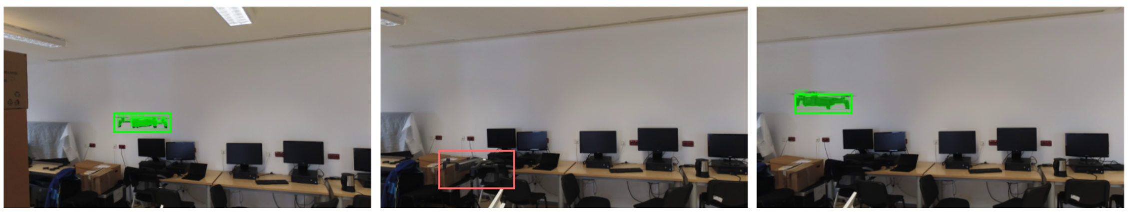

Figure 1: Success (left, right) and failure (middle) detection cases. An instance (middle case) where GroundingSAM misses the drone object. These cases were rare (1.4% of total frames) and manually corrected to maintain dataset quality.

Given the substantial labeling workload, we employed an automated labeling technique to generate pseudo-labels. We utilized GroundingSAM [19] for instance segmentation annotation, from which bounding boxes were subsequently extracted. By providing a simple ontology with class descriptions such as *"drone": "flying vehicle"*, the system achieved high annotation accuracy. Manual verification revealed that only approximately 200 frames out of 14,400 required correction, demonstrating a **98.6%** automated labeling accuracy.

Paper	Dataset	No of videos	No of images
Ours	**Ours(only drone footage)**	**4**	**14,400**
Real-Time Flying OD with YOLOv8	Dataset 1	No videos	15,064
Real-Time Flying OD with YOLOv8	Dataset 2	No videos	11,998
Dogfight	NPS-Drones	14	70,250
Dogfight	FL-Drones	50	38,948

Table 1: Dataset comparison showing our indoor drone dataset maintains comparable size to datasets used in [18] for YOLOv8 fine-tuning. A key distinction is the temporal information preserved in our video-based dataset. While smaller than [2] datasets, our high-quality annotations address potential labeling inconsistencies present in larger datasets.

3.2 Object Detection

To enable real-time onboard perception, we evaluate multiple YOLOv8 variants (Nano, Small, Large), following the methodology in [18]. Our student detector is trained using pseudo-labels generated by a frozen vision-language model (GroundingSAM), which annotates high-resolution drone video frames with minimal human supervision. The dataset consists of 1,395 training, 133 validation, and 67 test frames, all resized to 640×640 resolution with standard data augmentations. Videos 3 and 4 serve as temporal holdout sets to test generalization. Additionally, we include 1,115 frames of toy cars across three classes to evaluate cross-domain robustness using the same automated labeling pipeline. We assess model performance across precision, recall, mAP50-95, and inference latency to highlight trade-offs between accuracy and real-time feasibility (Table 2). YOLOv8-nano trained for 25 epochs achieves the best balance: 0.97 precision, 0.97 recall, 0.79 mAP50-95, and a low 4.7 ms inference time. While YOLOv8-small slightly improves precision (0.98), it incurs a 49% increase in latency (7.0 ms), making it less suitable for latency-constrained UAV deployment.

Figure 2: Qualitative examples of YOLOv8-based target following. UAVs track various targets including ground vehicles and aerial drones across diverse conditions—indoor classrooms, constrained spaces, and outdoor scenes—demonstrating cross-environment robustness. Demo video at `https://youtu.be/67IEGqvVRYI`

Model	Epochs	Precision ↑	Recall ↑	mAP50-95 ↑	Inference Time ↓ (ms)
YoloV8-nano	5	0.95	0.92	0.67	4.7
	10	0.97	0.97	0.72	4.7
	25	0.97	**0.97**	**0.79**	**4.7**
YoloV8-small	5	0.93	0.93	0.62	6.9
	10	0.95	0.94	0.72	7.0
	25	**0.98**	**0.97**	0.78	7.0
YoloV8-large	5	0.86	0.83	0.56	29.6
	10	0.94	0.93	0.69	32.4
	25	0.97	0.95	0.77	31.3

Table 2: Comparative evaluation of YOLOv8 architectures across training epochs, demonstrating the precision-efficiency trade-off critical for onboard deployment. Bold values indicate selected configuration optimizing for real-time constraints.

3.3 Visual Servoing Control

Traditional IBVS requires full 6-DoF interaction matrices and explicit depth estimation, which is often unreliable and numerically unstable in monocular UAV setups [4]. We propose a reduced 4-DoF controller tailored to UAV dynamics that eliminates depth dependency while preserving tracking accuracy.

Design Rationale: UAVs exhibit coupled pitch/roll with translational motion, allowing velocity control to be decoupled and reduced to 4 DoF.

Feature Extraction: From YOLO-based segmentation, we fit ellipses to targets, extracting center (x_c, y_c), axes (a, b), and orientation θ. A 14D feature vector $\mathbf{s}$ is built from seven key ellipse points: center, vertical offsets, and axis endpoints.

Interaction Matrix: For each point $\mathbf{p}_i = [x_i, y_i]^T$ relative to image center (c_x, c_y) and dimensions (W, H):

$$L_{\text{eff}}^{(i)} = \begin{pmatrix} 1 & 0 & \frac{x_i - c_x}{W} & -(y_i - c_y) \\ 0 & -1 & \frac{y_i - c_y}{H} & (x_i - c_x) \end{pmatrix} \tag{1}$$

Stacking over all seven points yields $L_{\text{eff}} \in \mathbb{R}^{14\times4}$.

Control Law: Velocity is computed as:

$$\mathbf{v}_{\text{eff}} = -\lambda L_{\text{eff}}^{\dagger}(\mathbf{s} - \mathbf{s}^*) \tag{2}$$

where λ is the control gain and $L_{\text{eff}}^{\dagger}$ is the pseudoinverse.

Limitations: The controller assumes fixed flight altitude, with accuracy degrading under height variation or partial occlusion. Continuous target visibility and sufficient texture are required for ellipse fitting.

4 Experimental Analysis

4.1 YoloV8 - GradCAM

To validate detection reliability, we employed GradCAM [21, 22] to visualize YOLOv8's attention patterns across different scenarios.

Attention Pattern Analysis: In mixed human-drone scenarios (Figure 3), the model shows strong bias toward human subjects, potentially compromising drone detection. Drone-only scenarios (Figure 4) exhibit distributed attention across multiple UAV targets with reduced environmental interference.

Figure 3: GradCAM showing human-biased attention

Figure 4: GradCAM demonstrating distributed drone attention in UAV-focused scenarios

Implications for Deployment: The observed attention bias suggests potential performance degradation in human-populated environments. This finding validates our dataset design focusing on indoor drone-to-drone scenarios, where human interference is minimized. The distributed attention pattern in drone-centric scenarios indicates robust multi-target tracking capabilities essential for autonomous operation.

4.2 Control Strategy

We evaluated our modified IBVS controller against baseline PID across straight-line and circular tracking scenarios using detection heatmaps and control command analysis.

Performance Comparison: Table 3 shows IBVS achieves superior tracking with 89% (straight-line) and 92% (circular) centering scores versus PID's 68% and 71%. IBVS demonstrates higher control authority but increased command variability.

Scenario	**Controller**	**Range v_x/v_y/ω_z Cmd (m/s)**	**Centering Score**	**Smoothness**
Straight Line	PID	[0, 0]/[-10, 30]/[-4, 4]	0.68	High
Straight Line	IBVS	[-4, 4]/[-10, 40]/[0, 25]	**0.89**	Medium
Circular	PID	[0, 0]/[-10, 30]/[-5, 15]	0.71	Low
Circular	IBVS	[-10, 20]/[-20, 30]/[0, 9]	**0.92**	High

Table 3: Controller performance comparison. Centering Score represents the fraction of time the target remained within the central 30% of the image frame. Command Smoothness is qualitatively assessed from standard deviation analysis.

The depth-decoupled IBVS approach shows clear advantages in tracking accuracy and adaptability to complex trajectories. IBVS exhibits particularly improved performance during high-curvature segments where PID shows erratic behavior. The responsiveness-smoothness trade-off suggests potential for gain scheduling optimization.

5 Conclusion and Discussion

We presented a self-supervised framework for autonomous indoor UAV tracking that combines scalable dataset generation with depth-free visual servoing control. Leveraging a frozen vision-language model (GroundingSAM), we achieved 98.6% labeling accuracy across a high-resolution indoor drone dataset, significantly reducing manual annotation effort. Our evaluation of YOLOv8 variants highlights the trade-off between accuracy and inference latency. YOLOv8-nano strikes the best balance for real-time deployment (0.97 precision/recall at 4.7 ms), while larger models offer marginal gains at a cost to onboard performance. GradCAM analysis revealed attention bias in human-populated scenes, reinforcing our focus on drone-to-drone indoor scenarios. The proposed depth-decoupled IBVS controller enables stable target following using only 2D visual features. Compared to a PID baseline, it improves centering performance (89–92% vs. 68–71%), especially in complex trajectories. However, limitations remain: our controller assumes a fixed flight altitude, and detection reliability degrades under occlusion or low contrast, occasionally leading to tracking loss. Despite these constraints, our approach demonstrates a reproducible and scalable pipeline for markerless indoor UAV tracking. The integration of automated annotation and depth-independent control enables robust operation in GPS-denied environments, with potential applications in warehouse automation, surveillance, and multi-agent indoor robotics.

References

[1] Luis Arreola, A Montes De Oca, Alejandro Flores, J Sanchez, and Gerardo Flores. Improvement in the uav position estimation with low-cost gps, ins and vision-based system: Application to a quadrotor uav. In *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1248–1254. IEEE, 2018.

[2] Muhammad Waseem Ashraf, Waqas Sultani, and Mubarak Shah. Dogfight: Detecting drones from drones videos, 2021.

[3] Sandeep Banik, Jinrae Kim, Naira Hovakimyan, Luca Carlone, John Thomas, and Nancy G Leveson. Integrating vision systems and stpa for robust landing and take-off in vtol aircraft. In *AIAA SCITECH 2025 Forum*, page 1324, 2025.

[4] Peter I Corke et al. *Visual Control of Robots: high-performance visual servoing*. Research Studies Press Taunton, UK, 1996.

[5] Anton Domini Sta Cruz, Jayson Osayan, Luis Rafael Moreno, and Rob Christian Caduyac. Performance of aruco detector in aruco marker detection using non-gps drone. In *Advancing Sustainable Science and Technology for a Resilient Future*, pages 63–67. CRC Press, 2024.

[6] Nicolas Guenard, Tarek Hamel, and Robert Mahony. A practical visual servo control for an unmanned aerial vehicle. *IEEE Transactions on Robotics*, 24(2):331–340, 2008.

[7] Seth Hutchinson, Gregory D Hager, and Peter I Corke. A tutorial on visual servo control. *IEEE transactions on robotics and automation*, 12(5):651–670, 1996.

[8] Michail Kalaitzakis, Brennan Cain, Sabrina Carroll, Anand Ambrosi, Camden Whitehead, and Nikolaos Vitzilaios. Fiducial markers for pose estimation: Overview, applications and experimental comparison of the artag, apriltag, aruco and stag markers. *Journal of Intelligent & Robotic Systems*, 101:1–26, 2021.

[9] Alexandros Kouris and Christos-Savvas Bouganis. Learning to fly by myself: A self-supervised cnn-based approach for autonomous navigation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018.

[10] Yogesh Kumar, Bassam Pervez Shamsi, Sayan Basu Roy, and Sujit PB. Tracking a planar target using image-based visual servoing technique. *IEEE Transactions on Intelligent Vehicles*, 9(3):4362–4372, 2024.

[11] Daewon Lee, Hyon Lim, H Jin Kim, Youdan Kim, and Kie Jeong Seong. Adaptive image-based visual servoing for an underactuated quadrotor system. *Journal of Guidance, Control, and Dynamics*, 35(4):1335–1353, 2012.

[12] E. Marchand, F. Spindler, and F. Chaumette. Visp for visual servoing: a generic software platform with a wide class of robot control skills. *IEEE Robotics and Automation Magazine*, 12(4):40–52, December 2005.

[13] Eduard Mráz, Jozef Rodina, and Andrej Babinec. Using fiducial markers to improve localization of a drone. In *2020 23rd International Symposium on Measurement and Control in Robotics (ISMCR)*, pages 1–5. IEEE, 2020.

[14] Tobias Nägeli, Lukas Meier, Alexander Domahidi, Javier Alonso-Mora, and Otmar Hilliges. Real-time planning for automated multi-view drone cinematography. *ACM Transactions on Graphics (TOG)*, 36(4):1–10, 2017.

[15] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*, pages 3400–3407. IEEE, 2011.

[16] Chao Qin, Qiuyu Yu, HS Helson Go, and Hugh H-T Liu. Perception-aware image-based visual servoing of aggressive quadrotor uavs. *IEEE/ASME Transactions on Mechatronics*, 28(4):2020–2028, 2023.

[17] Sairam VC Rebbapragada, Pranoy Panda, and Vineeth N Balasubramanian. C2fdrone: Coarse-to-fine drone-to-drone detection using vision transformer networks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6627–6633. IEEE, 2024.

[18] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8, 2023.

[19] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. URL `https://arxiv.org/abs/2401.14159`.

[20] Tushar Sangam, Ishan Rajendrakumar Dave, Waqas Sultani, and Mubarak Shah. Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos, 2023.

[21] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

[23] Jingtao Sun, Yaonan Wang, and Danwei Wang. Towards real-world aerial vision guidance with categorical 6d pose tracker. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[24] Jian Tang, Zhiyuan Deng, Hao Zhang, Tao Jiang, Chenyang Wang, and Zhiwu Ke. Control barrier function-based quadrotor interception control using visual servoing. In *2024 IEEE International Conference on Unmanned Systems (ICUS)*, pages 319–324. IEEE, 2024.

[25] Celine Teuliere, Laurent Eck, and Eric Marchand. Chasing a moving target from a flying uav. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4929–4934. IEEE, 2011.

[26] Lukas Wawrla, Omid Maghazei, and Torbjørn Netland. Applications of drones in warehouse operations. *Whitepaper. ETH Zurich, D-MTEC*, 212:1–12, 2019.

[27] Kun Yang and Quan Quan. An autonomous intercept drone with image-based visual servo. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2230–2236. IEEE, 2020.

[28] Kun Yang, Chenggang Bai, Zhikun She, and Quan Quan. High-speed interception multicopter control by image-based visual servoing. *IEEE Transactions on Control Systems Technology*, 2024.

[29] Yongwei Zhang, Yangguang Yu, Shengde Jia, and Xiangke Wang. Autonomous landing on ground target of uav by using image-based visual servo control. In *2017 36th Chinese Control Conference (CCC)*, pages 11204–11209. IEEE, 2017.