Stereo YOLO UAV Localization and Tracking Enabling Autonomous Sensor Deployment on Critical Infrastructure

Mark Zheng*, Md Asifuzzaman Khan†, Joud N. Satme‡, Korebami O. Adebajo§, Ryan Yount¶, and Austin R.J. Downey‖

Unmanned Aerial Vehicle (UAV)-based structural health monitoring (SHM) systems have gained significant traction in recent years. These systems typically collect data using onboard UAV sensors (such as cameras) or through the placement of physical sensors on the structure itself (such as strain gauges). While UAVs could be manually piloted to deploy these sensors, this approach is limited in hazardous or hard-to-reach areas. To address this, we present the use of object-tracking machine learning algorithms to support end-to-end autonomous control systems for sensor placement. The algorithm is trained on multiple images of the UAV within the flight environment to determine its relative position in three-dimensional space. These positional coordinates provide critical data on the UAV's location, guiding it toward the desired sensing point on the structure. Our system employs two externally placed cameras, oriented 90 degrees apart, to enable a full 3D assessment of the UAV's surroundings during flight. These cameras feed visual data to a processor, which analyzes the imagery and outputs coordinates for the UAV. These are then compared to the known coordinates of the target structure. Based on this comparison, motion commands for accurate alignment and docking can be calculated. This approach enhances the UAV's spatial awareness during flight, enabling safer and more reliable sensor deployment. The setup also reduces the risk of collisions, facilitates operations in hazardous environments, and enables sensor placement in previously inaccessible areas. Moreover, the algorithm is adaptable and capable of tracking various objects for broader applications in sensor delivery and visual navigation. The mean squared error of the model showed no further improvement after 96 of 196 epochs of training, with a training box loss of 0.770 square pixels and a validation box loss of 1.227 square pixels. Overall model summary after training and validation consists of 218 layers and 25,840,339 parameters for the UAV. Validation errors stabilized at the end of the training process, while training errors showed continual improvements. Average model error during data collection was 5.850 pixels for X-direction movement, 3.744 pixels for Y-direction movement, and 3.534 pixels for Z-direction movement of the UAV.

I. Nomenclature

$YOLO$	=	You Only Look Once
UAV	=	Unmanned Aerial Vehicle
SHM	=	Structural Health Monitoring
IOU	=	Intersection Over Union
$Confusion\ matrix$	=	Proportion table of predicted and true class instances
$Confidence$	=	Likelihood of a true positive

*Undergraduate Research Assistant, Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA.
†Graduate Research Assistant, Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA.
‡Graduate Research Assistant, Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA.
§Undergraduate Research Assistant, Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA.
¶Graduate Research Assistant, Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA.
‖Associate Professor, AIAA Member, Department of Mechanical Engineering, Department of Civil and Environmental Engineering, University of South Carolina, Columbia, SC 29208, USA, Contact Author austindowney@sc.edu

II. Introduction

Effective structural health monitoring (SHM) is essential to ensure the safety of aging infrastructures. Many critical structures continue operating beyond their intended design life, often without modern sensing systems. Retrofitting them with new sensors is both expensive and logistically challenging, particularly in remote or hazardous locations such as bridges and high-voltage towers. To address these limitations, recent research [1] has increasingly focused on unmanned aerial vehicle (UAV)-based SHM systems as a flexible and cost-effective alternative. These systems primarily rely on vision-based technologies integrated within the UAV, such as onboard cameras, sensors, and navigation systems [2], which assist not only in piloting but also in data collection from the structure.

Stereo vision has emerged as a promising method in UAV-based SHM for capturing structural movement. This approach uses high-contrast features or speckle patterns (dot arrays) on a structure [3]. Stereo cameras—typically mounted on each end of the UAV system and calibrated to the structure's known positions; can record the displacement of specific points [3]. This data enables the calculation of structural parameters such as stress and strain. However, long-term evaluations are often better served by sensor deployment to the structure itself. UAVs can aid in this task by transporting and placing sensors using machine learning (ML) algorithms. This method is particularly valuable in inaccessible areas and provides more detailed analysis than stereo vision alone, which can be compromised by flight instability, visibility issues, or environmental interference. Reliable sensor placement thus depends on robust autonomous flight path planning [4]. Identifying points of interest and efficient routes along the structural environment are key to delivering optimal results within the UAV battery life.

Machine learning-enhanced computer vision has become central to UAV navigation in SHM applications. During operations, ML-based object detection allows UAVs to avoid obstacles [5]. These systems use mapped flight paths supported by ultrasonic beacons and fiducial markers [6]. Beacons emit signals that help determine distances, while camera-based ML algorithms detect fiducial markers [5]. The combination of ultrasonic and visual data feeds into a machine-learning model and tunes a proportional-integral-derivative (PID) controller, enabling accurate UAV navigation. This setup can also facilitate geotagging structural damage in areas where GPS coverage is limited [6].

Aerodynamic effects and limited camera perspectives complicate UAV navigation in confined structural environments. Ceiling and ground effects are two common phenomena when UAVs operate near overhead or underlying surfaces, posing serious stability challenges [7]. With the ceiling effect, UAVs experience a significant thrust increase when flying near these environments [8]. These problems are especially relevant when UAVs try to dock sensors to the underside of a bridge or while navigating in tight spaces. Narrow access points may trigger both effects simultaneously, increasing the risk of crashes. Compounding the issue is the restricted field of view of onboard cameras, which limits environmental awareness.

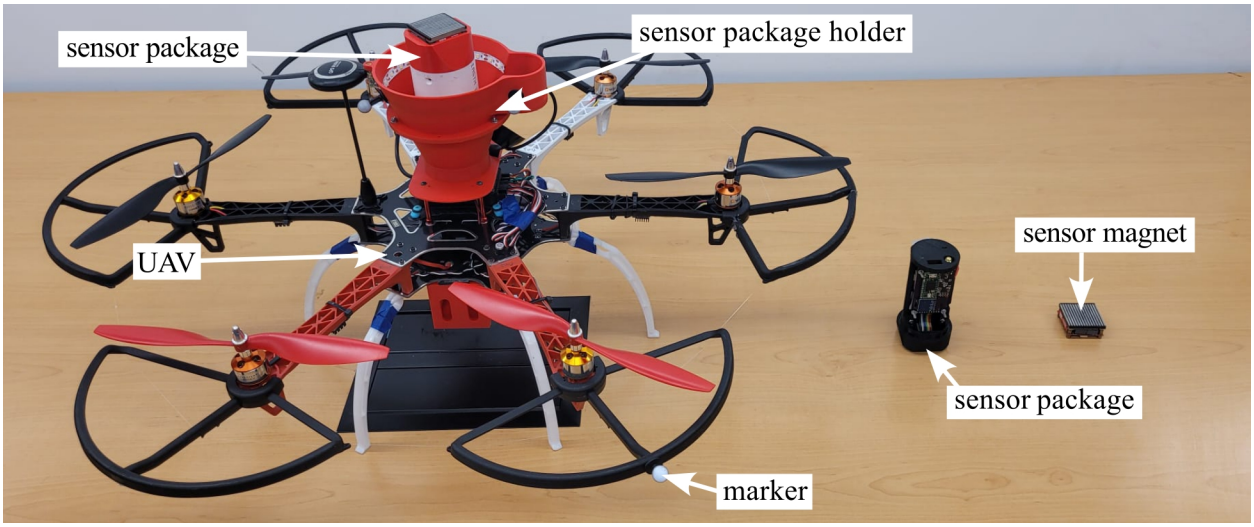

Fig. 1 UAV equipped with reflective silver marker and sensor package deployment system.

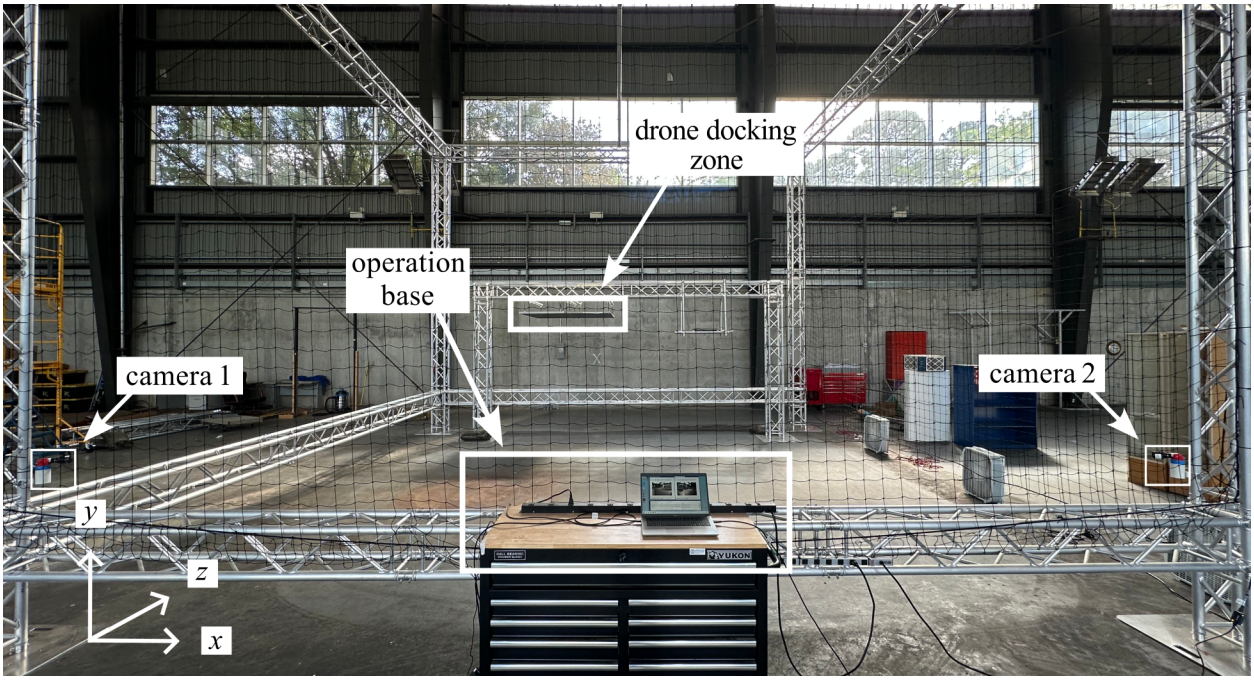

Fig. 2 Complete UAV system used in experimental setups and testing sites, with the YOLO algorithm running on the videos recorded by the cameras to analyze the output.

To address these challenges, an external vision-assisted navigation system that covers the entire operational area can significantly enhance flight path safety. Andrean et al. proposed a system employing a dual-camera setup positioned 90 degrees apart to detect a UAV marked with a specific color [9]. The image processing algorithm tracks this color to estimate the UAV's position. Although this approach is computationally efficient, it requires the UAV to maintain a predefined color, and the presence of similar colors in the background can reduce detection accuracy. Rifqi et al. introduced another system utilizing a ceiling-mounted single camera to track a UAV through the YOLO machine learning algorithm [10]. The UAV's depth is estimated from the bounding box around the detected region. However, this method faces a generalization issue, as the depth estimation remains accurate only for UAV models similar to those used during the YOLO model's training phase.

This paper introduces an external stereo vision system for enhancing UAV localization during sensor docking. By installing stereo cameras within the structural environment and applying ML algorithms, we achieve precise tracking of UAV coordinates as sensors are placed on a surface modeled after that of an under-bridge sensor placement. Unlike beacon- or marker-dependent systems, this vision approach relies solely on visual features of the UAV and docking zone, eliminating the need for structural modifications. The two main contributions of this work are: (1) the development of a strategic placement method for stereo vision cameras that enables accurate three-dimensional UAV tracking, and (2) the design of an ML algorithm that processes camera inputs to estimate the UAV's position relative to the target docking zone.

III. Methodology

To establish a suitable testing environment for the UAV system, a metallic structure, a UAV, and a sensor package were constructed. The UAV is equipped with a reflective silver marker and incorporates an electro-permanent-magnet-based sensor deployment mechanism [11, 12], as shown in Figure 1. The metallic docking structure consists of three large beams: two vertical beams supporting a horizontal beam, where the sensor is intended to be attached (Figure 2). This structure is referred to as the docking structure in Figure 2. Relevant data, code, and artifacts related to this paper are available through a public repository [13].

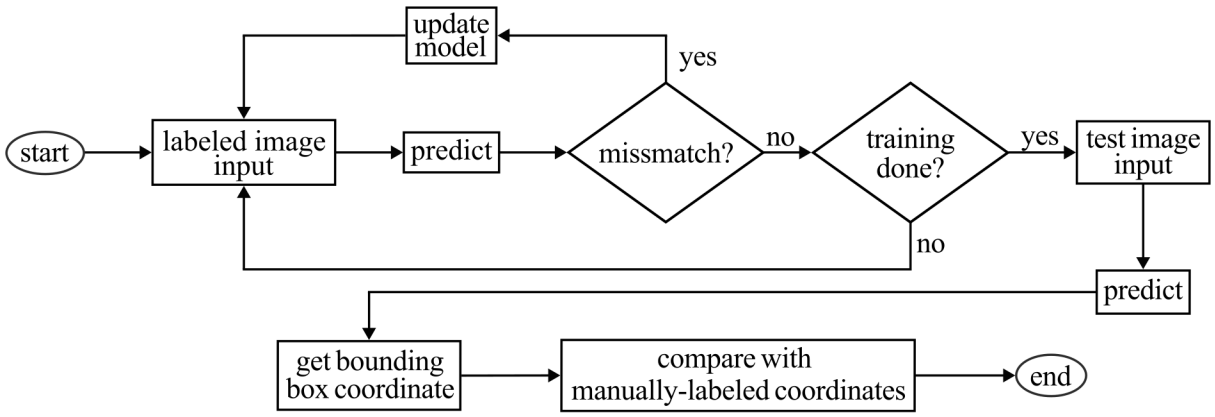

Fig. 3 Training and detection method explained in a flow chart.

A. You Only Look Once (YOLO)

A trained You Only Look Once (YOLO)-based object tracking algorithm was developed to visually track the UAV using only camera feeds. This algorithm runs on a laptop connected to two external cameras oriented toward the flight area, as illustrated in Figures 2. The YOLO model was trained using 64 hand-labeled images of the UAV in the given environment. Each training image was annotated with a bounding box around the UAV's position in the frame. By comparing labeled positions to predicted outputs, the algorithm adjusted itself using the statistical trends between the observed errors and the true locations, refining its accuracy over iterations.

After training, the model's output was validated using a control group consisting of 6 uniformly time-distributed frames extracted from a 77-second UAV flight video. Each frame was manually annotated to mark the UAV's position based on the reflective silver marker, located at the UAV's center, making it a reliable reference point for position determination. The coordinate system was defined as follows: the X-axis was derived from the horizontal axis of the left camera, the Z-axis from the horizontal axis of the right camera, and the Y-axis from the average of the vertical axes of both cameras. The Y-axis represented UAV height, the X-axis corresponded to lateral movement (left/right relative to the docking structure), and the Z-axis represented the distance from the camera/control station.

Coordinates were extracted from the bounding boxes generated by the customized YOLO algorithm for the 7 selected frames. In object-tracking ML, a bounding box indicates the region where the algorithm predicts the object exists with confidence above a set threshold (25% in this experiment). For symmetrical objects like the UAV, the center of the bounding box offers a reliable estimate of object location. This entire process is depicted in Figure 3.

To ensure consistent tracking performance, several parameters were fine-tuned: confidence threshold, intersection over union (IOU), and frame rate. The confidence threshold filters out detections with low probability, removing likely false positives. IOU helps eliminate duplicate tracking results by rejecting overlapping detections beyond a specified threshold. The frame rate influences both tracking quality and computational load, depending on system deployment conditions. Specifically, the confidence threshold was set at 25%, IOU at 0.1, and frame rate matched the camera's input footage at 60 frames per second (FPS). IOU values range from 0 (no overlap) to 1 (complete overlap), while frame rate represents the number of image frames processed per second.

B. Determining YOLO Accuracy

To evaluate the accuracy of the YOLO algorithm, an additional UAV test flight was conducted. For the test flight, YOLO-predicted UAV positions were compared to manually marked ground truth coordinates across 277 uniformly time-distributed frames of the 55-second flight. Both datasets were plotted on the same 3D coordinate system (XYZ plot) using consistent axis definitions. Linear interpolation was applied between each pair of adjacent frames to generate continuous trajectories for both the hand-labeled and algorithm-generated positions. The deviation between these two trajectories represents the error of the YOLO algorithm. The objective is for the UAV to dock the sensor package

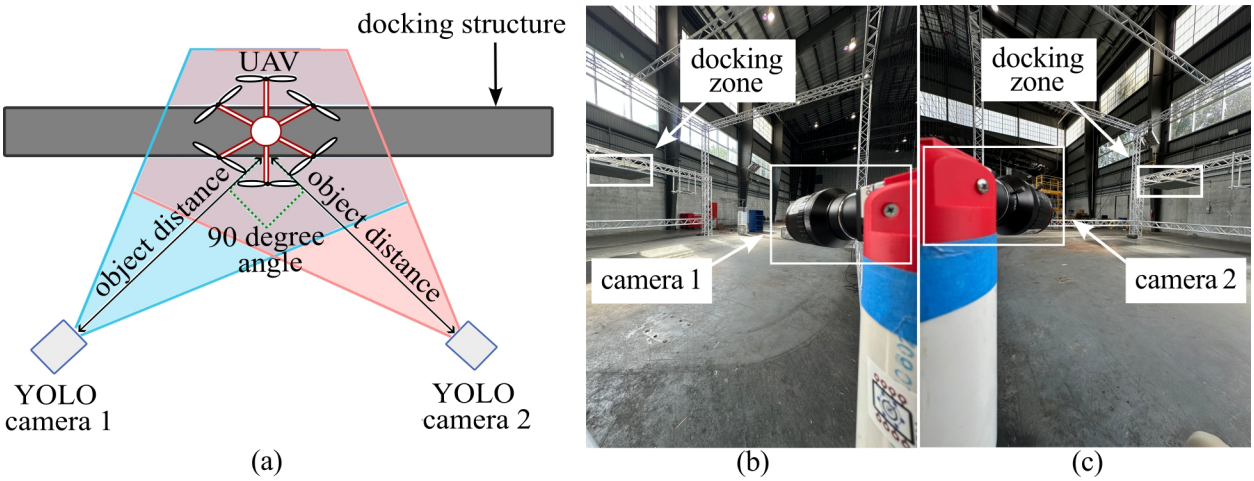

Fig. 4 Camera positions in the experimental set-up, showing: (a) the angle depiction between the cameras; (b) the position of the left camera in the two-camera set-up, and; (c) the position of the right camera in the two-camera set-up, when facing towards the structure and back wall.

onto the structure, with the YOLO-based tracking algorithm accurately identifying the UAV's position during this docking process. A successful result would show a close match between the 3D trajectories of the manually marked and ML-predicted coordinates. If accurate enough, these predicted coordinates can be translated into directional commands, guiding the UAV to dock the sensor in the designated zone while avoiding collisions with surrounding structures.

C. Technical Aspects of the Experiment

The experiment was conducted using two color cameras (FLIR BFS-U3-63S4C-C) operating at approximately 60 FPS. Each camera was equipped with an adjustable (varifocal) lens set to a 2.8 mm focal length to achieve a wide-angle view. One camera was positioned at 45° angle to the left of the central axis, facing the structural docking zone. The other camera was placed at 45° angle to the right, forming a 90° angle between their fields of view, as illustrated in Figure 4. The right and left camera was positioned 7.065 m and 5.980 m from the center of the docking zone, respectively, and a measuring tape was used to ensure accurate placement. Each camera was angled directly horizontal from the ground plane and mounted 1.055 m above ground level.

The structure targeted for UAV docking created an open area measuring 152.3 × 152.3 cm in height and width. It was composed of three square posts, with each post constructed from four metal beams joined at the corners. Two vertical posts supported a third horizontal post spanning across the top. The central point of this top post served as the primary docking location for the UAV and was properly marked in both camera views for YOLO-based processing. Both cameras recorded simultaneously during each UAV flight to ensure synchronized visual data collection.

IV. Results

After conducting multiple experimental UAV flights, sufficient data were collected to demonstrate that the YOLO algorithm was consistently generating bounding boxes around the UAV for the majority of each flight, as illustrated in Figure 5. This performance was verified using YOLO's annotation software, which automatically produced video outputs with bounding box overlays on the UAV. A confidence threshold of 25% effectively filtered out nearly all extraneous detections while still providing multiple accurate coordinate data points per second. Likewise, an Intersection over Union (IOU) value of 0.1 successfully minimized the occurrence of duplicate bounding boxes over the same object.

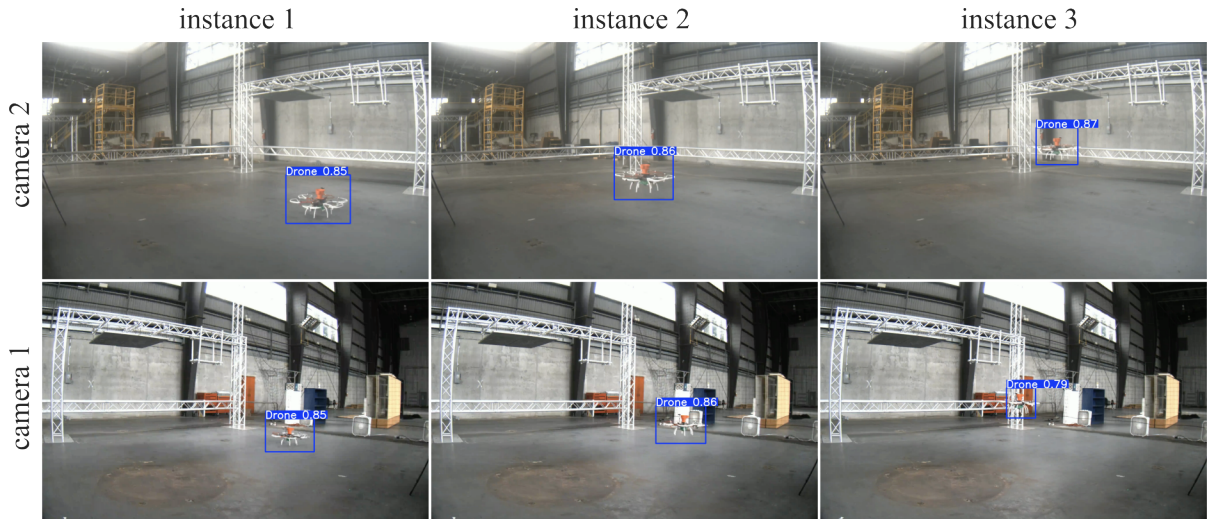

Fig. 5 YOLO object-tracking output on the drone in a variety of scenarios, with these photos used to train and analyze the object detection patterns exhibited by YOLO.

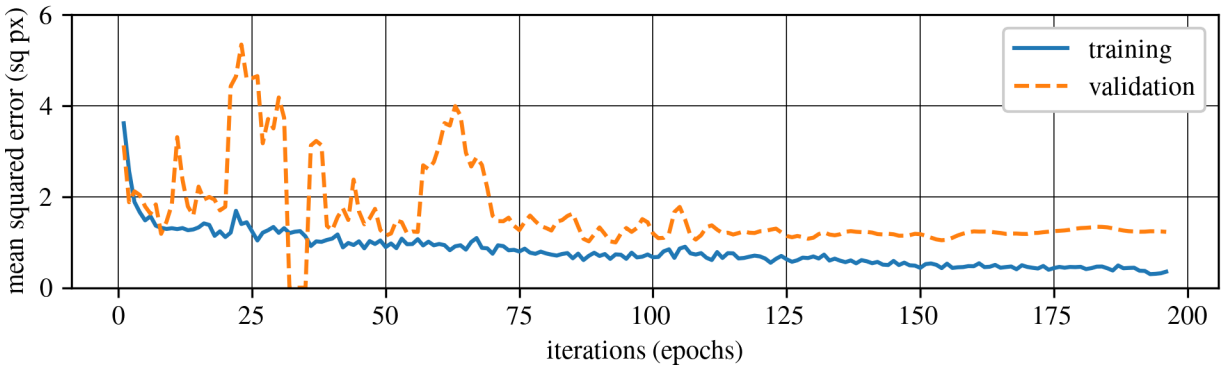

Fig. 6 Learning curve of training results from YOLO algorithm over a series of 196 epochs on multiple images gathered of the drone as it flies toward the structure. Lowest consistent mean squared error at 96 epochs.

Mean squared error of the YOLO algorithm bounding box in tracking the experimental UAV's movement to its docking zone were evaluated using the plot shown in Figure 6. The data for this plot was derived from training the YOLO model on 64 hand-labeled images (training) of the UAV within the experimental environment. The algorithm's predictions were then compared against a set of 6 randomly selected hand-labeled images (validation) to assess accuracy. The mean squared error of the model showed no further improvement after 96 of 196 epochs of training, with a training box loss of 0.770 square pixels and a validation box loss of 1.227 square pixels. Overall model summary after training and validation consists of 218 layers and 25,840,339 parameters for the UAV. Validation errors stabilized at the end of the training process, while training errors showed continual improvements.

Figure 7 shows YOLO predicted coordinates and manually marked reference coordinates with respect to time, over a 55 s UAV flight in the testing area. Furthermore, the error is also plotted as the reference coordinates subtracted from the predicted coordinates. Parameters were calculated for all three cardinal directions of UAV movement. Average error for the X, Y, and Z directions was -5.850, 3.744, and 3.534 pixels, respectively. The 3D plot of the UAV's flight path generated by the YOLO algorithm, shown in Figure 8, closely matched the manually marked coordinates. The largest deviations between the two datasets occurred during periods of drone rotation. While the YOLO algorithm tracked the drone's center consistently, the manually marked coordinates, based on the reflective silver marker, varied slightly due to changes in the marker's position during rotation.

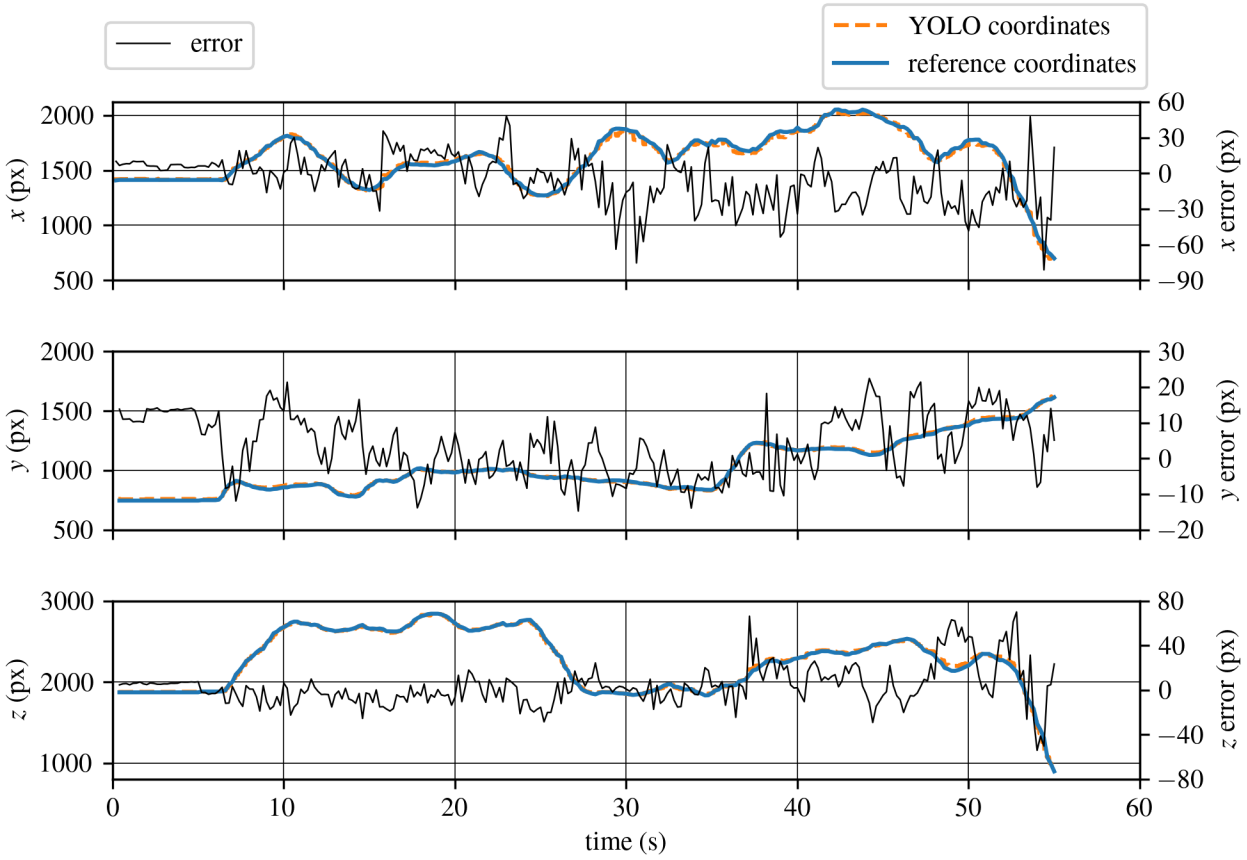

Fig. 7 Comparison between YOLO generated coordinates and hand-marked coordinates with respect to time, along each axis of UAV movement.

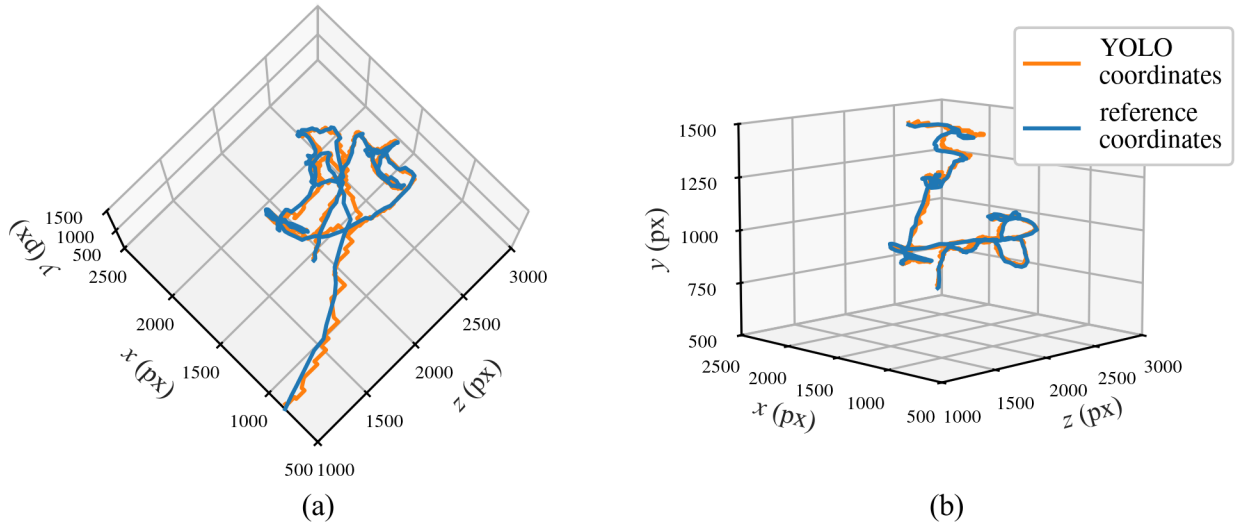

Fig. 8 3D position plot of manually marked coordinates and coordinates retrieved from YOLO, showing: (a) the ground level view of the flight path, and; (b) the top level view of the flight path in the UAV cage.

V. Conclusions

This research demonstrates that machine-learning-based tracking systems can be readily developed to provide accurate positional data for guiding an unmanned aerial vehicle to dock sensor packages onto a predetermined zone on a structure. Using only two cameras combined with a You Only Look Once (YOLO) algorithm, the system generated three-dimensional coordinates comparable to manually labeled data. These coordinates were updated multiple times per second, offering sufficient temporal resolution for the low-speed UAV maneuvers required during sensor docking.

Model performance was also good from a numerical perspective. The mean squared error, in square pixels (sq px), was calculated through training of the YOLO model in UAV bounding box creation over a set of 64 training images and 6 validation images. No mean squared error improvement was observed after 96 epochs, with a training box loss of 0.770 sq px and a validation box loss of 1.227 sq px. This is in the context of 3840 x 2160 px imaging data resized to 640 x 640 px squares for the training process. Overall model summary after training and validation consists of 218 layers (hyperparameters and architectural components of the model) and 25,840,339 parameters (data interpretation and detection conditions) for the UAV. Validation errors were unstable towards the beginning of the training process, but eventually stabilized as the model was able to consistently detect the UAV. Training errors showed continual improvements as it is more of an additive process in building a stronger model. All of this is consistent with expected trends in model creation and performance.

Future work will focus on continuously refining the YOLO algorithm to achieve even closer alignment with manually labeled coordinates. These improved positional outputs could then be integrated into a Proportional-Integral-Derivative (PID) controller to enable fully automated UAV operations in a wider range of applications.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research (AFOSR) through award no. FA9550-21-1-0083. This work is also partly supported by the National Science Foundation (NSF) grant numbers CCF - 1956071, CMMI - 2152896, CCF-2234921 and, CPS - 2237696. Additional funding for this work comes from the Office of Naval Research through the award number N000142412727. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force, the National Science Foundation, or the United States Navy.

References

[1] Carroll, S., Satme, J., Alkharusi, S., Vitzilaios, N., Downey, A., and Rizos, D., "Drone-Based Vibration Monitoring and Assessment of Structures," *Applied Sciences*, Vol. 11, No. 18, 2021, p. 8560. https://doi.org/10.3390/app11188560.

[2] Fayyad, T. M., Taylor, S., Feng, K., and Hui, F. K. P., "A scientometric analysis of drone-based structural health monitoring and new technologies," *Advances in Structural Engineering*, Vol. 28, No. 1, 2024, pp. 122–144. https://doi.org/10.1177/13694332241255734.

[3] Reagan, D., Sabato, A., and Niezrecki, C., "Feasibility of using digital image correlation for unmanned aerial vehicle structural health monitoring of bridges," *Structural Health Monitoring*, Vol. 17, No. 5, 2017, pp. 1056–1072. https://doi.org/10.1177/1475921717735326.

[4] Panigati, T., Zini, M., Striccoli, D., Giordano, P. F., Tonelli, D., Limongelli, M. P., and Zonta, D., "Drone-based bridge inspections: Current practices and future directions," *Automation in Construction*, Vol. 173, 2025, p. 106101. https://doi.org/https://doi.org/10.1016/j.autcon.2025.106101.

[5] Waqas, A., Kang, D., and Cha, Y.-J., "Deep learning-based obstacle-avoiding autonomous UAVs with fiducial marker-based localization for structural health monitoring," *Structural Health Monitoring*, Vol. 23, No. 2, 2023, pp. 971–990. https://doi.org/10.1177/14759217231177314.

[6] Kang, D., and Cha, Y., "Autonomous UAVs for Structural Health Monitoring Using Deep Learning and an Ultrasonic Beacon System with Geo-Tagging," *Computer-Aided Civil and Infrastructure Engineering*, Vol. 33, No. 10, 2018, pp. 885–902. https://doi.org/10.1111/mice.12375.

[7] Conyers, S. A., Rutherford, M. J., and Valavanis, K. P., "An Empirical Evaluation of Ceiling Effect for Small-Scale Rotorcraft," *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 2018. https://doi.org/10.1109/icuas.2018.8453469.

[8] Satme, J. N., Yount, R., Schwartz, S., Downey, A. R. J., and Ling, S., "Analyzing the Ceiling Effect on UAV Thrust Variations Through Computational and Experimental Studies," 2025, p. V006T10A028. https://doi.org/10.1115/DETC2025-169020.

[9] Himawan, R. W., Baylon, P. B. A., Sembiring, J., and Jenie, Y. I., "Development of an Indoor Visual-Based Monocular Positioning System for Multirotor UAV," *2023 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES)*, 2023, pp. 1–7. https://doi.org/10.1109/ICARES60489.2023.10329792.

[10] Andrean, S. Y., Joelianto, E., Widyotriatmo, A., and Adiprawita, W., "Low cost vision-based 3D localization system for indoor unmanned aerial vehicles," *2013 International Conference on Robotics, Biomimetics, Intelligent Computational Systems*, 2013, pp. 237–241. https://doi.org/10.1109/ROBIONETICS.2013.6743611.

[11] Satme, J. N., Yount, R., Goujevskii, N., Jannazzo, L., and Downey, A. R. J., "Sensor Package Deployment and Recovery Cone with Integrated Video Streaming for Rapid Structural Health Monitoring," *ASME 2024 Conference on Smart Materials, Adaptive Structures and Intelligent Systems*, American Society of Mechanical Engineers, 2024. https://doi.org/10.1115/smasis2024-140435.

[12] Smith, C., Satme, J., Martin, J., Downey, A. R., Vitzilaios, N., and Imran, J., "UAV rapidly-deployable stage sensor with electro-permanent magnet docking mechanism for flood monitoring in undersampled watersheds," *HardwareX*, Vol. 12, 2022, p. e00325. https://doi.org/10.1016/j.ohx.2022.e00325.

[13] ARTS-Lab, "Paper 2026 Stereo YOLO UAV Localization and Tracking," GitHub, 2026. URL https://github.com/ARTS-Laboratory/Paper-2026-Stereo-YOLO-UAV-Localization-and-Tracking.