Mitigating Occlusion and Re-Identification Challenges in UAV Object Tracking

Tai Ho[1], Trong-An Bui[2,*], Pei-Jun Lee[3], Hsin-Piao Lin[2],
Thanh-Thoai Le[1], Daniel Selva[4], Hai Tran[1]

[1]Ho Chi Minh University of Education, HCMC, Vietnam
[2]National Taipei University of Technology, Taipei, Taiwan (R.O.C)
[3]National Taiwan University of Science and Technology, Taipei, Taiwan (R.O.C)
[4]Texas A&M University, TX, U.S.A

***Abstract*—Object tracking in UAV imagery is challenging due to occlusions, scale variations, and frequent object disappearance. YOLO-based detection lacks temporal consistency, resulting in ID switches and tracking failures when objects reappear after occlusion. To address this, this study proposes an enhanced tracking framework integrating YOLO for object detection, DeepSORT for identity assignment, and an occlusion handling module leveraging motion prediction and feature embeddings for re-identification. Evaluations on the UAVDT dataset demonstrate that the proposed approach significantly reduces ID switches and improves tracking stability compared to standalone YOLO tracking. The results highlight its effectiveness in real-world UAV applications, making it a robust solution for aerial surveillance and monitoring.**

***Index Terms*—UAV object tracking, occlusion handling, motion prediction, YOLO, DeepSORT, re-identification**

I. Introduction

The analysis of drone and UAV images has gained significant attention in various fields, including defense, agriculture, disaster management, and urban planning. However, tracking moving objects in UAV-captured images remains a challenging task due to factors such as rapid viewpoint changes, occlusions, varying flight altitudes, and object scale variations. Although the YOLO-based model [1] is highly efficient for object detection, it struggles with temporal consistency, often resulting in identity changes and difficulties in maintaining continuous object trajectories. Figure 1 illustrates tracking

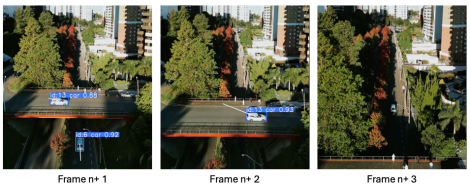

Fig. 1: Object tracking by the YOLO-based model

inconsistencies in UAV imagery across multiple frames. The car labeled as *id:13* maintains its identity, indicating stable tracking. However, *id:6* appears only in the first frame and disappears in subsequent frames, suggesting a loss of tracking. This dropout may result from occlusions, detection failures, or changes in object scale. In this particular case, disconnection occurs when the object is obscured by another object (a bridge), which prevents YOLO from detecting it in subsequent frames.

This work was supported by the Ministry of Education, Taiwan, *2024 UAAT-TAMUS Integrated Program: Interdisciplinary Capacity Building for the New Space Generation*.

*Corresponding author: Trong-An Bui trongan93@ntut.edu.tw.

Furthermore, the absence of a robust re-identification mechanism within YOLO exacerbates tracking discontinuities. When an object is momentarily undetected, the model fails to reassign its previous identity upon reappearance, leading to tracking loss or identity switching. These limitations are particularly significant in UAV-based tracking, where objects frequently move in and out of the camera's field of view due to dynamic flight paths and varying perspectives. To address these challenges, this paper introduces advanced techniques aimed at enhancing object tracking in UAV imagery by mitigating occlusion effects and ensuring temporal consistency.

II. Proposed Methodology

Figure 2 presents the proposed framework for UAV-based object tracking, *addressing the challenges of occlusion and re-identification*. The proposed approach integrates *YOLO-based* [2] object detection with *DeepSORT tracking* [3] to achieve persistent tracking of moving objects in UAV imagery. The tracking pipeline consists of three main components: (1) *YOLO-based object detection*, (2) *DeepSORT for identity assignment and tracking*, and (3) *Occlusion handling and re-identification*. The pipeline begins with *UAV-captured video*

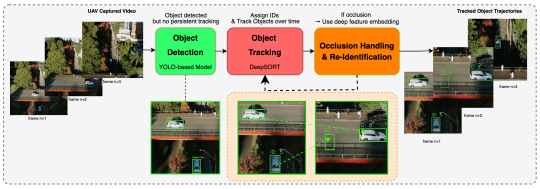

Fig. 2: Proposed Methodology Architecture

frames, where YOLO detects objects of interest and generates bounding boxes. These bounding boxes are defined based on the object's center coordinates $(x_{\text{center}}, y_{\text{center}})$ and dimensions (width, height), as formulated in Eq. (1).

$$x_{\min} = x_{\text{center}} - \frac{\text{width}}{2}, \quad y_{\min} = y_{\text{center}} - \frac{\text{height}}{2}, \qquad (1)$$
$$x_{\max} = x_{\text{center}} + \frac{\text{width}}{2}, \quad y_{\max} = y_{\text{center}} + \frac{\text{height}}{2}.$$

where: $x_{\min}, y_{\min}$ represent the coordinates of the top-left corner of the bounding box, while $x_{\max}, y_{\max}$ define the bottom-right corner. These four values encapsulate the spatial extent of each detected object within the video frame.

Once the bounding boxes are computed, YOLO assigns a *confidence score* to each detection, indicating the probability that the object belongs to a specific class (vehicle, pedestrian). However, since YOLO processes each frame independently, it

does not maintain object identities over time. To address this, *DeepSORT* is incorporated to assign unique IDs to detected objects and track them over time.

To represent the detected object numerically, the object region is first *cropped* from the input UAV image:

$$I_B = I[x_{\min} : x_{\max}, y_{\min} : y_{\max}] \tag{2}$$

where I_B represents the cropped image region containing the detected object. The cropped image is then *resized* to a fixed input size (h, w):

$$I'_B = \text{Resize}(I_B, h, w) \tag{3}$$

To extract object appearance features, the resized image is processed through a deep convolutional neural network (CNN) feature extractor $f(\cdot)$:

$$d = f(I'_B) \in \mathbb{R}^n \tag{4}$$

where d is the *appearance feature vector* representing the detected object, $f(\cdot)$ is a pre-trained CNN model (e.g., ResNet, MobileNet) used for feature extraction, and n is the *feature vector dimensionality*.

To associate new detections with previously tracked objects, DeepSORT compares feature vectors using the cosine similarity metric:

$$\mathcal{S}_{\cos}(d_1, d_2) = \frac{\mathbf{d_1} \cdot \mathbf{d_2}}{|\mathbf{d_1}||\mathbf{d_2}|} \tag{5}$$

where $\mathbf{d_1}, \mathbf{d_2}$ are the feature vectors of two objects.

A *higher similarity score* indicates that the two detections likely correspond to the same object. If $\mathcal{S}_{\cos}(d_1, d_2)$ exceeds a predefined threshold, the object retains its track ID; otherwise, a new ID is assigned. By integrating YOLO-based bounding boxes (Eq.1) with DeepSORT's appearance-based association (Eq.5), proposed framework ensures robust and consistent object tracking across UAV video frames.

Despite the improvements brought by DeepSORT, tracking inconsistencies may still arise due to *occlusion*, where objects momentarily disappear from the UAV's field of view due to obstacles (bridges, trees, or other vehicles). To mitigate these issues and prevent identity switches, this research introduce an **Occlusion Handling and Re-Identification Module**, which enhances object tracking by leveraging deep feature embeddings and motion-based predictions.

To handle occlusions, Kalman filter is used to predict the future state of the objects. The predicted state is given by:

$$x_k = A \cdot x_{k-1} + B \cdot w_k + w_k, \tag{6}$$

where x_k represents the predicted state (position and velocity), A is the state transition matrix, B is the control matrix, and w_k denotes process noise. The output from the cosine distance calculation, which helps track the objects' positions, is used as the input for the Kalman prediction.

In the correction step, real-time measurements are integrated to refine the predicted state, as expressed in Eq. (7):

$$z_k = H \cdot x_k + v_k, \tag{7}$$

where z_k represents the real-time measurement, H is the measurement matrix, and v_k is the measurement noise. The Kalman filter continuously updates the object's position, ensuring robust tracking even when temporary occlusions occur.

By integrating motion-based predictions with re-identification using feature embeddings, the proposed method significantly reduces identity switches and tracking inconsistencies. Finally, the output video is generated, displaying each tracked object with a unique ID and a smoothed trajectory, ensuring reliable and stable tracking across frames, even in complex UAV-based scenarios.

III. Experiments

This research utilizes the UAVDT dataset, which consists of 10 hours of UAV-captured video [4]. It features 100 sequences, totaling 80,000 frames at 30 fps, with a resolution of 1080x540 pixels. To evaluate tracking performance, the proposed architecture is compared with YOLO-based tracking on the UAVDT dataset. YOLO detects objects independently in each frame but struggles with occlusions and objects exiting the field of view, leading to missed detections and frequent ID switches. As shown in Figure 3, from (a) to (b), an object becomes occluded, and by (c), it is assigned a new ID due to tracking failure.

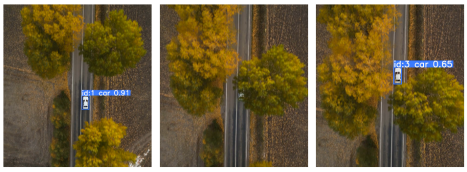

Fig. 3: Tracking by YOLO (a) Frame n+1, (b) Frame n+2, (c) Frame n+3

In contrast, the proposed system improves tracking consistency. In Figure 4, from (a) to (b), the object remains occluded but retains its ID due to motion prediction, and by (c), the correct ID is maintained.

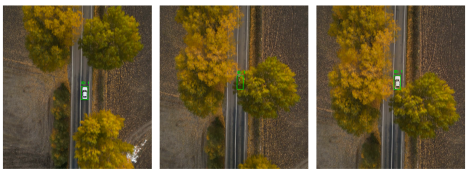

Fig. 4: Tracking by Proposed (a) Frame n+1, (b) Frame n+2, (c) Frame n+3

The results demonstrate that the proposed architecture enhances tracking stability by **reducing ID switches** and **improving occlusion handling**. The system provides a more accurate and robust tracking solution compared to standalone YOLO tracking, ensuring better performance in real-world UAV applications.

References

[1] P.-H. Hsu, P.-J. Lee, T.-A. Bui, and Y.-S. Chou, "Yolo-spd: Tiny objects localization on remote sensing based on you only look once and space-to-depth convolution," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 2024, pp. 1–3.

[2] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," *arXiv preprint arXiv:2305.09972*, 2023.

[3] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *CoRR*, vol. abs/1703.07402, 2017. [Online]. Available: http://arxiv.org/abs/1703.07402

[4] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.