Multi-Object Tracking in UAV Videos: A YOLOv11 Fusion Method for Detection and Segmentation Optimization

Mahmoud Ahmed[1*], Naser El-Sheimy[2]

[1] Dept. of Electrical and Software Engineering, University of Calgary, 2500 University Dr NW, Calgary, Alberta T2N 1N4 Canada – (mahmoud.ahmed2@ucalgary.ca)

[2] Dept. of Geomatics Engineering, University of Calgary, 2500 University Dr NW, Calgary, Alberta T2N 1N4 Canada – (elsheimy@ucalgary.ca)

KEY WORDS: Multi-Object Tracking, YOLOv11, Fusion, Segmentation, Object Detection.

ABSTRACT:

The rapid evolution of deep learning has significantly advanced multi-object tracking (MOT) in UAV-based remote sensing applications. However, accurately detecting and tracking objects of varying sizes in complex UAV-captured environments remains a challenge. This research introduces a novel fusion-based approach that leverages YOLOv11, a state-of-the-art object detection framework, to enhance MOT performance on the VisDrone UAV dataset. The proposed method integrates two YOLOv11 configurations: detection mode, paired with the Bot-SORT tracker, optimized for large objects to ensure high precision and localization accuracy, and segmentation mode, combined with the Byte-Track tracker, designed to effectively detect and track smaller, less prominent objects. By fusing the outputs of these configurations, the approach ensures comprehensive object coverage across different size ranges, thereby improving both detection and tracking accuracy while enhancing segmentation performance. This method addresses critical limitations in existing models, such as low recall for small objects and imprecise localization for larger ones, which are particularly challenging in UAV datasets due to varying altitudes, occlusions, and dynamic backgrounds. The fusion strategy employs Intersection over Union (IoU)-based matching, weighted bounding box fusion, and confidence thresholding to enhance tracking reliability and accuracy. Experimental evaluations on the VisDrone dataset, using motion tracking metrics and the F1 score for detection and segmentation, demonstrate significant performance improvements across multiple UAV videos. The results show that the fused approach outperforms individual configurations while maintaining consistent object identity tracking over time. This research contributes to UAV-based remote sensing by providing a scalable and efficient MOT framework, making it particularly valuable for applications such as surveillance, traffic monitoring, and disaster response, where precise object localization and tracking are crucial.

1. INTRODUCTION

Multi-object tracking and segmentation (MOTS) have gained significant attention in UAV-based applications such as surveillance, traffic monitoring, and disaster response. Despite advancements in deep learning for object detection and segmentation, tracking multiple objects remains a challenge, especially in dynamic environments with occlusions, scale variations, and motion blur. Traditional tracking-by-detection methods, which rely on bounding box annotations, often struggle with crowded scenes and overlapping objects. To address these limitations, pixel-wise segmentation techniques are being explored for improved accuracy and object identity preservation(Voigtlaender et al., 2019).

Video Object Segmentation (VOS) techniques aim to track objects at the pixel level, but existing datasets lack sufficient object diversity and fail to handle identity switches a critical issue in MOT. Recent research has focused on integrating object detection models with trackers like SORT, DeepSORT, and JDE, which enhance tracking accuracy but still face challenges such as occlusions and re-identification(Wang et al., 2020).

MOT methods can be broadly categorized into three main types(Li et al., 2025): detection-based, single-object tracking (SOT)-based, and segmentation-based, as illustrated in Fig.1 . Detection-based MOT, which includes approaches such as tracking-by-detection, Joint detection and tracking, and transformer-enhanced detection-based MOT, links object detections across consecutive frames, offering flexibility in various applications but facing challenges such as occlusions and identity switches. In contrast, SOT-based MOT tracks individual objects continuously throughout a sequence without relying on external detections in each frame, making it particularly effective for scenarios requiring long-term tracking of a specific target. Meanwhile, segmentation-based MOT provides pixel-level accuracy, reducing issues like bounding box overlaps and occlusion-induced errors. However, this approach is computationally intensive, limiting its feasibility for real-time applications. These categories highlight the diverse strategies employed in MOT, each with its advantages and trade-offs, depending on the specific tracking requirements and computational constraints.

For UAV-based MOT, detection tasks often rely on object detection networks, categorized into two-stage and single-stage models. Two-stage networks, like Faster R-CNN, offer high precision but slower inference times. In contrast, single-stage networks like SSD and YOLO prioritize speed and efficiency, making them well-suited for real-time UAV applications(Cao et al., 2021). The introduction of yolov11, with its improved detection capabilities, is particularly beneficial for UAV-based MOT, as it enhances accuracy while maintaining high processing speeds(Lui et al., 2024a).

However, challenges persist, such as occlusions, scale variability, and appearance changes due to lighting and perspective shifts. Real-time constraints also require lightweight architecture. Recent advancements in data association techniques, including IoU-based matching and appearance-based re-identification, have improved MOT performance but have not fully addressed these challenges(Zhu et al., 2022).

This research presents a novel fusion-based method that combines YOLOv11 with Bot-SORT and ByteTrack to enhance UAV-based multi-object tracking and segmentation. By leveraging Detection Mode for large objects and Segmentation Mode for smaller ones, the approach improves tracking accuracy, object identity preservation, and segmentation performance in dynamic environments. Key contributions include:

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-M-8-2025-1-2025 | © Author(s) 2025. CC BY 4.0 License.

- Fusion of YOLOv11 Configurations: Integrates Detection Mode (with Bot-SORT) and Segmentation Mode (with ByteTrack) to effectively track both large and small objects.
- Improved Tracking and Identity Preservation: Fusing output enhances accuracy, identity retention, and segmentation, addressing challenges like occlusion and dynamic backgrounds.
- Advanced Fusion Strategy: Utilizes IoU-based matching, weighted bounding box fusion, and confidence thresholding to boost reliability and localization.
- Efficient Accuracy-Time Trade-off: Balances processing speed and precision, enabling real-time tracking for UAV applications.
- Scalability and Real-time Performance: Designed for real-world UAV tasks such as surveillance, traffic monitoring, and disaster response.

Overall, the method offers a robust, scalable solution for real-time object detection, tracking, and segmentation in complex UAV environments

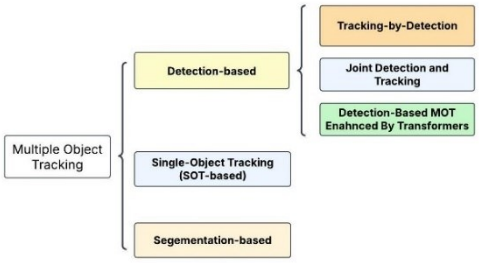

Multiple Object Tracking
Detection-based
Tracking-by-Detection
Joint Detection and Tracking
Detection-Based MOT Enahnced By Transformers
Single-Object Tracking (SOT-based)
Segementation-based

Figure 1. The classification of MOT considers its integrated core technologies.

2. MATERIALS AND METHODS

The proposed methodology enhances object detection and tracking in UAV-based applications like traffic monitoring, environmental surveillance, and disaster management, where challenges such as scale variation, occlusion, and dynamic backgrounds are common. To address these, we introduce a dual-mode framework that integrates YOLOv11 with Bot-SORT for large-object tracking and ByteTrack for small-object tracking, as shown in Fig. 2.

YOLOv11 operates in detection mode for large objects and segmentation mode for small ones. Unlike traditional bounding boxes, segmentation masks offer precise, per-pixel boundaries—crucial for accurately detecting small or occluded objects in cluttered scenes. This improves both detection and tracking reliability in dynamic environments.

The segmentation mode combined with ByteTrack enhances small object tracking, reducing missed detections often seen with bounding boxes. Treating detection and tracking as a unified task leads to a more accurate and coherent solution for real-time UAV applications.

Post-detection, Bot-SORT handles large objects while ByteTrack manages smaller ones. Their outputs are fused using IoU-based matching, confidence thresholding, and weighted averaging, optimizing accuracy and reducing false associations.

The following sections detail YOLOv11’s dual-mode capabilities, the tracking mechanisms of Bot-SORT and ByteTrack, and the multi-scale fusion strategy. This flexible, efficient approach is well-suited for real-time UAV surveillance in complex, dynamic environments.

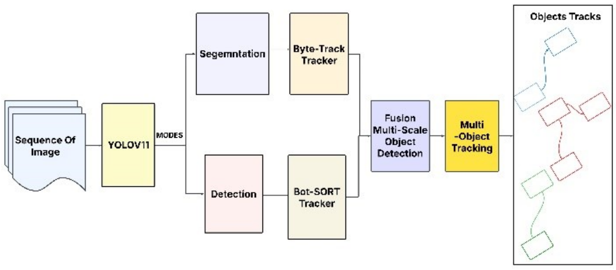


Figure 2. The Proposed Framework.

2.1 YOLOv11 Architecture

YOLOv11 marks a major advancement in real-time object detection for UAV-based remote sensing, addressing key challenges like scale variation, occlusion, and dynamic backgrounds. Building on previous YOLO versions, it introduces architectural improvements that boost both speed and accuracy in complex environments.

As illustrated in Fig. 3three main components (Khanam and Hussain, 2024): The backbone uses the new C3k2 block, an efficient variant of the CSP bottleneck, along with SPPF and C2PSA modules to enhance spatial attention and feature extraction. The neck integrates multi-scale features, replacing C2f with C3k2 and incorporating C2PSA to improve focus on small and occluded objects. The head refines predictions using C3k2 and CBS layers, improving localization and classification.

YOLOv11 also features Multi-Scale Adaptive Feature Fusion (MAFF) for handling varying object sizes, Dynamic Head Attention (DHA) for better focus on small targets, and an Enhanced Anchor-Free Prediction system for more accurate localization without predefined anchors. Advanced training strategies like Mosaic v3 and MixUp further improve generalization.

Beyond object detection, YOLOv11 supports a range of tasks including instance segmentation, image classification, pose estimation, oriented object detection (OBB), and real-time tracking. Its versatility makes it ideal for UAV-based applications such as surveillance, traffic monitoring, and environmental analysis, as well as broader domains like medical imaging, e-commerce, and robotics.

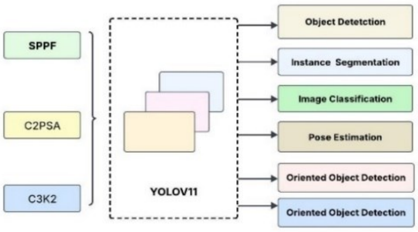


Figure 3. Key architectural modules in YOLO11.

2.2 Detection Mode with Bot-SORT Tracker

This section focuses on the detection of large objects using YOLOv11's detection mode, paired with the Bot-SORT tracker. Detection Mode is optimized for identifying larger objects within UAV imagery, providing precise bounding boxes to mark their locations. After detection, Bot-SORT is employed to track these objects over time, ensuring consistent and accurate monitoring of their trajectories. This combination is designed to handle the complexities of multi-object tracking, particularly in environments where large objects move in close proximity or interact with each other. Integrating detection mode with Bot-SORT ensures high detection accuracy and robust tracking performance for large objects in dynamic, real-time UAV applications.

2.2.1 YOLOv11 in Detection Mode

YOLOv11's Detection Mode excels at detecting large objects, such as vehicles, buildings, and ships in UAV imagery. Key architectural improvements, including a CSPDarknet backbone with lightweight convolutional blocks and attention-based feature refinement, enhance feature extraction while maintaining efficiency. The integration of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) ensures robust feature propagation across scales, improving large-object detection and spatial detail preservation (Alif, 2024).

To improve object localization, YOLOv11 uses IoU-Aware Object Localization with the Wise IoU (WIoU) loss function. This refines bounding box regression by incorporating penalties for aspect ratio and distance via Complete IoU (CIOU) and Distance IoU (DIOU).

$$L_{WIoU} = 1 - \left(\lambda_1 IoU + \lambda_2 CIOU + \text{ DIOU }^2\right) \tag{1}$$

where λ_1 and λ_2 control the balance between IoU, aspect ratio, and distance penalties. This loss function improves localization accuracy, particularly for objects with elongated shapes or varying orientations, ensuring more precise bounding box predictions.

2.2.2 Bot-SORT for Tracking Large Objects:

For improved multi-object tracking, YOLOv11's detection mode integrates with the Bot-SORT tracker (Aharon et al., 2022), enhancing tracking accuracy through advanced motion prediction and object association. Kalman Filtering is used to predict object motion across frames using the state transition model.

$$x_{k+1} = Fx_k + w_k \tag{2}$$

where F represents the state transition matrix, and w_k accounts for process noise. This predictive modeling ensures smoother object trajectory estimation, particularly in dynamic UAV scenes.

To handle occlusions and reappearances, Re-Identification (ReID) Embeddings are used to distinguish objects across frames by extracting discriminative features. Additionally, IoU and Motion-Based Association with the Hungarian algorithm optimize object matching across frames which is defined as :

$$M = \arg \max_{A} \sum_{(i,j)\in A} IoU\left(b_i, b_j\right) \tag{3}$$

where b_i and b_j represent bounding boxes in consecutive frames. By prioritizing IoU-based associations, the tracker ensures robust object matching, reducin ↓ entity switches and improving long-term tracking accuracy.

By combining YOLOv11's large-object detection with the Bot-SORT tracker's stability, this configuration ensures accurate object identification and trajectory estimation in UAV imagery. It enables precise tracking of large objects in complex aerial scenes, ideal for traffic monitoring, infrastructure assessment, and maritime surveillance.

2.3 Segmentation Mode with Byte-Track Tracker

This section covers small object detection using YOLOv11's segmentation mode and the Byte-Track tracker. Segmentation mode provides per-pixel boundaries, ideal for detecting small objects in cluttered or occluded environments. Byte-Track ensures reliable tracking, even with occlusions or interactions, enhancing accuracy and robustness in real-time UAV applications.

2.3.1 YOLOv11 in Segmentation Mode

Detecting small objects like pedestrians, drones, and wildlife in UAV imagery is challenging due to low resolution and background clutter. YOLOv11 improves small-object detection with advanced segmentation and a Dual Attention Mechanism (DAM), which combines spatial and channel attention to enhance feature focus and reduce noise, boosting accuracy for small or occluded targets.

YOLOv11 incorporates High-Resolution Feature Refinement using dilated convolutions to expand the receptive field while preserving spatial details, enhancing segmentation accuracy (Ultralytics, n.d.). It also introduces a boundary-aware loss for improved foreground-background separation and precise small-object localization.

$$L_{Seg} = L_{CE} + \gamma L_{\text{Boundary}} \tag{4}$$

where L_{CE} represents the cross-entropy loss for classification, and L_{Boundary} penalizes errors in object boundaries. The inclusion of boundary-aware penalties ensures that small objects are well-defined, reducing misclassifications caused by blending with the background. These enhancements collectively enable yolov11 to achieve superior segmentation performance, particularly for small and complex objects in UAV-based remote sensing applications.

2.3.2 Byte-Track for Tracking Small Objects

To improve small-object tracking, YOLOv11's segmentation mode is combined with the ByteTrack tracker, which refines detection-to-tracklet associations using a two-stage matching process (Zhang et al., 2022). High-confidence detections (τ_h), are matched using IoU and the Hungarian algorithm, while low-confidence ones ($\tau_l < \tau < \tau_h$), e retained and refined based on motion consistency. The match score is computed as:

$$\text{Match Score } = \alpha \cdot IoU + (1-\alpha) \cdot \text{ Feature Similarity} \quad (5)$$

Here, α balances IoU-based matching and feature similarity, ensuring both spatial proximity and visual cues are considered. By retaining and refining low-confidence detections, the integration of YOLOv11's segmentation mode with ByteTrack improves small-object tracking, ensuring accurate localization and trajectory estimation in dynamic UAV environments.

2.4 Fusion Strategy for Multi-Scale Object Detection

To achieve robust multi-scale object detection and tracking, a weighted fusion strategy is employed to integrate bounding box predictions from detection mode (Bot-SORT) and segmentation mode (Byte-Track). This approach ensures that both large and small objects are accurately detected and tracked while addressing inconsistencies between the outputs of the two trackers. The fusion process is carried out in several steps, beginning with confidence thresholding. In this step, low-confidence detections are filtered out to reduce false positives. A detection is considered valid if its confidence score C_i meets the threshold

$$Ci \geq \tau c \quad (6)$$

where τc is a predefined confidence threshold, set to 0.5 in this implementation. If both detections have confidence scores below this threshold, they are discarded. Following confidence thresholding, Intersection over Union (IoU) matching is used to determine whether detections from both trackers correspond to the same object. The IoU between two bounding boxes B1 and B2is calculated as

$$\text{IoU } (B_1, B_2) = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} \quad (7)$$

where $|B_1 \cap B_2|$ represents the intersection area and $|B_1 \cup B_2|$ represents the union area of the bounding boxes. If IoU (B_1,B_2)<τ_i, where τ_i is the loU threshold, the detections are considered independent, and the bounding box with the higher confidence score is selected.

When two bounding boxes have sufficient loU overlap, we apply weighted bounding box fusion to compute a refined bounding box. The fused bounding box B_f is determined as

$$B_f = w_1 B_1 + w_2 B_2 \quad (8)$$

where w1 and w2 are the normalized fusion weights, determined based on the confidence scores C1 and C2 of the two bounding boxes:

$$w_1 = \frac{C_1}{C_1 + C_2} \quad (9)$$

$$w_2 = \frac{C_2}{C_1 + C_2} \quad (10)$$

This ensures that more reliable detections contribute more significantly to the final bounding box's position and size. The fused bounding box parameters, which include the top-left corner coordinates $x_f$and y_f, as well as the width $w_f$and height h_f, are calculated as follows:

$$x_f = w_1 x_1 + w_2 x_2 \quad (11)$$

$$y_f = w_1 y_1 + w_2 y_2 \quad (12)$$

$$w_f = w_1 w_1 + w_2 w_2 \quad (13)$$

$$h_f = w_1 h_1 + w_2 h_2 \quad (14)$$

The confidence score for the fused bounding box C_f is calculated using a weighted average:

$$C_f = w_1 C_1 + w_2 C_2 \quad (15)$$

For the class ID, visibility, and unused attributes, the values from the detection with the higher confidence score are adopted, ensuring consistency across frames. If a bounding box appears in only one tracker, or if the IoU between two detections is below the threshold τi, the bounding box with the higher confidence score is retained. This ensures that objects missing by one tracker are still included in the final fused output, preserving complete detection across varying object scales. The final selection rule is given by:

$$B_f = \arg \max_{B_i} C_i, \text{ if IoU } (B_1, B_2) < \tau_i \quad (16)$$

This process guarantees that high-confidence detections are prioritized, while lower-confidence detections are integrated adaptively, ensuring robust and comprehensive detection and tracking across both large and small objects.

2.5 Material

The VisDrone dataset, developed by the AISKYEYE team at Tianjin University, is a comprehensive benchmark for drone-based computer vision tasks (Du et al., 2019). It comprises 288 video clips (261,908 frames) and 10,209 static images captured across 14 urban and rural cities in China. The dataset includes over 2.6 million manually annotated bounding boxes for diverse objects like pedestrians, vehicles, and bicycles, under varying weather and lighting conditions. It supports five tasks: object detection in images and videos, single- and multi-object tracking, and crowd counting. For our experiments, we used YOLOv11 pretrained on COCO, a dataset with 80 annotated object categories, to perform both detection and segmentation. Various configurations YOLOv11n for speed and YOLOv11x for accuracy were tested based on task needs(Lui et al., 2024b). These models effectively addressed both large-object detection and small-object segmentation challenges.

All experiments were run using PyTorch on a system with a 12th Gen Intel Core i7-12700H (2.70 GHz), 16 GB RAM, and a GeForce RTX 3070 GPU.

3. RESULT AND EVALUATION

Evaluation results for Video1 and Video2, based on F1-score, MOTP, and IDF1, highlight the superior performance of the Fusion method, which combines detection mode with Bot-SORT (L weight) for large-object accuracy and segmentation mode with Byte-Track (L weight) for enhanced small-object tracking. This configuration leverages the strengths of both approaches to deliver robust tracking across object scales. MOTP, which evaluates object localization accuracy(Andriluka et al., 2008), is given by:

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t} \quad (17)$$

where $d_{i,t}$ is the localization error of matched detections, and c_t is the number of correct matches at time t. Lastly, IDF1-score (Bernardin and Stiefelhagen, 2008), evaluates identity consistency across the tracking sequence and is computed as:

$$IDF1 = \frac{2 \times IDTP}{2 \times IDTP + IDFP + IDFN} \tag{18}$$

where IDTP are correctly assigned identities, IDFP are incorrect identity assignments, and IDFN are objects that lost their correct identity

Tracker	Mode	Weight	F1	MOTP	IDF1
Byte Tracker	Det	n	0.6484	0.5053	0.604
		s	0.7214	0.6083	0.6672
		m	0.7874	0.7	0.7542
		l	0.7912	0.7036	0.7235
		x	0.7693	0.6738	0.7186
	Seg	n	0.6727	0.5292	0.6166
		s	0.7634	0.6657	0.6622
		m	0.7897	0.7034	0.758
		l	0.794	0.7092	0.7293
		x	0.7705	0.6784	0.7236
Bot Sort	Det	n	0.712	0.586	0.653
		s	0.7691	0.6686	0.6891
		m	0.8314	0.7643	0.7782
		l	0.8461	0.7855	0.7549
		x	0.8171	0.7464	0.7475
	Seg	n	0.7252	0.6018	0.6499
		s	0.7896	0.7214	0.6601
		m	0.8165	0.7526	0.7663
		l	0.8279	0.7728	0.7431
		x	0.8008	0.737	0.7382
Our	Fusion		0.8636	0.8069	0.79

Table 1. Tracking Performance Metrics for Video 1.

Tracker	Mode	Weight	F1	MOTP	IDF1
Byte Tracker	Det	n	0.5224	0.3822	0.5121
		s	0.6712	0.5676	0.6376
		m	0.6498	0.5355	0.6194
		l	0.7239	0.632	0.6832
		x	0.7237	0.6556	0.6817
	Seg	n	0.5668	0.4327	0.551
		s	0.6296	0.5144	0.6025
		m	0.6544	0.5376	0.6218
		l	0.7313	0.6369	0.6884
		x	0.7287	0.658	0.6841
Bot Sort	Det	n	0.5382	0.4002	0.525
		s	0.6884	0.5908	0.6482
		m	0.6696	0.5652	0.6359
		l	0.7404	0.6628	0.6967
		x	0.7398	0.6846	0.6907
	Seg	n	0.5871	0.4597	0.5693
		s	0.641	0.5355	0.6115
		m	0.6668	0.5599	0.63
		l	0.7369	0.6577	0.6913
		x	0.7343	0.6778	0.6839
Our	Fusion		0.75	0.7221	0.6839

Table 2. Tracking Performance Metrics for Video 2.

The Fusion method, combining Bot-SORT Tracker in detection mode (L weight) and Byte-Track Tracker in segmentation mode (L weight), outperformed individual configurations for Video1 and Video2. For Video1, it achieved an F1-score of 0.863, MOTP of 0.8069, and IDF1 of 0.79, surpassing Bot-SORT (0.846) and Byte-Track (0.794) for small objects (Table 1). For Video2, the Fusion method scored an F1 of 0.75, MOTP of 0.7221, and IDF1 of 0.71, outperforming Bot-SORT (0.7404) and Byte-Track (0.7313) for small objects (Table 2). By dynamically switching modes, the Fusion method effectively handles occlusions, scale variations, and enhances tracking accuracy. Tests with five YOLOv11 configurations confirmed its superiority in multi-object tracking for surveillance and autonomous systems.

4. ANALYSIS AND DISCUSSION

The evaluation of processing times shows that the L weight configuration balances tracking precision and efficiency for both Byte-Track and Bot-SORT trackers. Bot-SORT in detection mode takes 339.4 ms/frame for Video1 as in Figure 4, while Byte-Track takes 338.4 ms/frame. In segmentation mode, Byte-Track requires 924.1 ms/frame for L weight in Video1. For Video2 as in Figure 5, Bot-SORT in detection mode takes 319.2 ms/frame, while Byte-Track in segmentation mode requires 929.1 ms/frame for X weight. The Fusion method, combining Bot-SORT for large objects and Byte-Track for small ones, ensures optimal performance with processing times of 631.75 ms/frame for Video1 and 401.4 ms/frame for Video2. This method is ideal for real-time tracking, especially for the VisDrone dataset, handling challenges like small objects and occlusions. Overall, the L weight configuration provides the best balance, making the Fusion method suitable for complex tracking tasks.

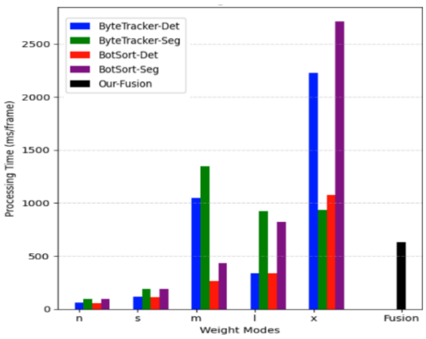


Figure 4. Processing Time Comparison for Video1.

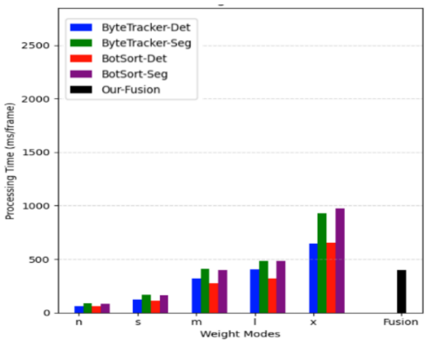
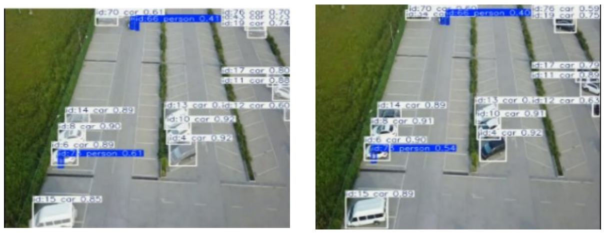


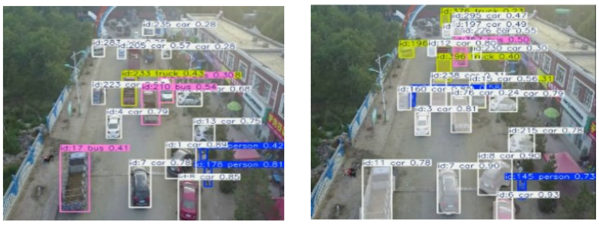
Figure 5. Processing Time Comparison for Video2.

Figures 6 and 7 showcase the fusion-based multi-object tracking method for UAV-captured urban environments. Figure 6 presents result for Video 1, and Figure 7 for Video 2, demonstrating the method's adaptability. Each figure includes two modes: (a) Detection Mode for large objects like vehicles and (b) Segmentation Mode for smaller objects like pedestrians. The fusion approach dynamically selects the optimal mode based on object size, improving tracking stability and reducing ID switches. By balancing tracking accuracy and processing speed, the method ensures robust performance, making it ideal for UAV-based surveillance, traffic monitoring, and intelligent transportation systems.



a) Tracking detection mode b) Tracking segmentation mode

Figure 6. Multi object tracking in video 1.



a) Tracking detection mode. b) Tracking segmentation mode

Figure 7. Multi object tracking in video 2.

5. CONCULSION

This study proposes a fusion-based approach to enhance object detection and tracking in UAV remote sensing. It leverages YOLOv11's segmentation mode for accurate small-object detection and detection mode for large-object precision, ensuring robustness across object sizes. Bot-SORT, used with detection mode, efficiently tracks large objects with low computational cost, while Byte-Track, paired with segmentation mode, improves tracking of small and occluded objects through refined detection associations. Their fusion balances accuracy and efficiency. A weighted strategy with confidence thresholding and IoU matching ensures reliable tracking. Experiments on the VisDrone dataset show the Fusion method outperforms individual configurations in F1-score, MOTP, and IDF1. In summary, this method offers a real-time, accurate, and efficient solution for UAV-based surveillance. Future work may focus on optimizing fusion strategies and expanding into other datasets and detection models.

6. REFERENCES

Aharon, N., Orfaig, R., Bobrovsky, B.-Z., 2022. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. https://doi.org/10.48550/arXiv.2206.14651

Alif, M.A.R., 2024. YOLOv11 for Vehicle Detection: Advancements, Performance, and Applications in Intelligent Transportation Systems. https://doi.org/10.48550/arXiv.2410.22898

Andriluka, M., Roth, S., Schiele, B., 2008. People-tracking-by-detection and people-detection-by-tracking, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. https://doi.org/10.1109/CVPR.2008.4587583

Bernardin, K., Stiefelhagen, R., 2008. Evaluating multiple object tracking performance: the CLEAR MOT metrics. J. Image Video Process. 2008, 1:1-1:10. https://doi.org/10.1155/2008/246309

Cao, Y., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., Hoi, S., Hu, Q., Liu, M., Cheng, C., Liu, F., Cao, G., Li, G., Wang, H., He, J., Wan, J., 2021. VisDrone-DET2021: The Vision Meets Drone Object detection Challenge Results, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Presented at the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 2847–2854. https://doi.org/10.1109/ICCVW54120.2021.00319

Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, Xinyao, Zhang, Yue, Bo, L., Shi, H., Zhu, R.,. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Presented at the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 213–226. https://doi.org/10.1109/ICCVW.2019.00030

Khanam, R., Hussain, M., 2024. YOLOv11: An Overview of the Key Architectural Enhancements. https://doi.org/10.48550/arXiv.2410.17725

Li, S., Ren, H., Xie, X., Cao, Y., 2025. A Review of Multi-Object Tracking in Recent Times. IET Computer Vision 19, e70010. https://doi.org/10.1049/cvi2.70010

Lui, M.H., Liu, H., Tang, Z., Yuan, H., Williams, D., Lee, D., Wong, K.C., Wang, Z., 2024a. An Adaptive YOLO11 Framework for the Localisation, Tracking, and Imaging of Small Aerial Targets Using a Pan–Tilt–Zoom Camera Network. Eng 5, 3488–3516. https://doi.org/10.3390/eng5040182

Lui, M.H., Liu, H., Tang, Z., Yuan, H., Williams, D., Lee, D., Wong, K.C., Wang, Z., 2024b. An Adaptive YOLO11 Framework for the Localisation, Tracking, and Imaging of Small Aerial Targets Using a Pan–Tilt–Zoom Camera Network. Eng 5, 3488–3516. https://doi.org/10.3390/eng5040182

Ultralytics, n.d. Instance Segmentation with Object Tracking [WWW Document]. URL https://docs.ultralytics.com/guides/instance-segmentation-and-tracking (accessed 3.24.25).

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. MOTS: Multi-Object Tracking and Segmentation. https://doi.org/10.48550/arXiv.1902.03604

Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S., 2020. Towards Real-Time Multi-Object Tracking. https://doi.org/10.48550/arXiv.1909.12605

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. https://doi.org/10.48550/arXiv.2110.06864

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H., 2022. Detection and Tracking Meet Drones Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence44,7380–7399.