Detection and Multiple Object Tracking of Military Vehicles Using YOLOv12 and Hybrid-SORT on Drone Video

Christopher Brian
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
13522106@std.stei.itb.ac.id

***Abstract*—In the ever-evolving nature of modern warfare, drones have become an essential equipment for reconnaissance and tracking of targets, including military vehicles. This generates visual data in the form of a large volume of videos which requires automatic detection and tracking to reduce the needs of human operators. However, real-life conditions met by drones used in war frequently encounter challenges, such as low resolution, low frame rate, as well as simultaneous motion of drone and military vehicles. These challenges can potentially lower the performance of military vehicles detection and tracking systems. This paper presents a detection and multiple object tracking (MOT) pipeline combining a fine-tuned YOLOv12-S detector with a Hybrid-SORT tracker and a fine-tuned OSNet-AIN re-identification module, specifically designed to operate under challenging conditions including low resolution (480p), low frame rate (6 FPS), and simultaneous motion of drone and military vehicles. To improve detection recall under these conditions, we employ multi-scale inference at resolutions of 1152p and 1280p with Weighted Boxes Fusion (WBF) for merging detections. Linear interpolation is then applied as post-processing to fill brief gaps in tracked objects. Evaluated on a drone-captured military vehicle dataset, our pipeline achieves a Higher Order Tracking Accuracy (HOTA) of 49.29% and an Identity F1-Score (IDF1) of 51.20% on the test set, surpassing the prior state-of-the-art result of 48.29% HOTA on the same data source. The full pipeline runs at 8.3 FPS on a single NVIDIA T4 GPU, meeting real-time requirements for the 6 FPS dataset.** *(Abstract)*

Keywords—computer vision; drone; Hybrid-SORT; military vehicle; multiple object tracking; object detection; YOLOv12

I. INTRODUCTION

The modern battlefield has been fundamentally transformed by the proliferation of unmanned aerial vehicles (UAVs), commonly known as drones. These platforms, initially developed for reconnaissance and intelligence gathering, have rapidly evolved into versatile tools for surveillance, target acquisition, electronic warfare, and precision strike. In recent armed conflicts, drones have demonstrated the capacity to conduct saturation attacks at costs far below those of conventional munitions, and their operational requirements can often be met through minimal modifications of commercial UAV platforms. This ubiquity has placed drone-based surveillance of military vehicles at the forefront of operational intelligence needs.

A key challenge arising from this operational reality is the sheer volume of video data generated by drone operations. Continuous aerial surveillance of ground forces produces streams of footage that are infeasible to analyze manually at the speed required for tactical decision-making. Automated detection and multiple object tracking (MOT) systems are therefore essential to reduce dependence on human operators and enable real-time situational awareness. However, the video data captured by low-altitude military drones is characterized by conditions that are severely hostile to conventional computer vision systems.

The principal technical challenges inherent to this domain include: (1) low spatial resolution, as drone video is frequently downsampled or received at lower resolutions due to bandwidth constraints and compression; (2) low temporal resolution, where frame rates are low enough to produce large inter-frame displacements that confound conventional motion-based association; (3) simultaneous motion of both the drone platform and the observed military vehicles, causing unstable camera perspectives and non-static backgrounds; (4) domain-specific visual characteristics of military vehicles, such as camouflage patterns that minimize contrast with natural terrain, making them visually difficult to distinguish from background clutter; and (5) object occlusion and brief disappearances from the camera field of view.

Existing MOT approaches address these challenges only partially. General-purpose models trained on standard pedestrian or traffic datasets do not transfer well to military vehicle domain. Research efforts targeting drone-based MOT, such as DroneMOT [1] and HDHNet [2], have demonstrated improvements on civilian UAV datasets like VisDrone2019 [3] and UAVDT [4], but do not account for the specific visual characteristics of military vehicles. Work directly addressing military vehicle tracking [5] has tackled low frame rate conditions but leaves simultaneous drone and vehicle motion largely unresolved. No prior published system simultaneously addresses all of these challenges in an integrated pipeline, while also retaining the capability to run inference real-time on a low-end GPU. This paper makes the following contributions: (1) An

end-to-end MOT pipeline tailored for military vehicle surveillance from drone video is presented, integrating a fine-tuned YOLOv12-S [6] detector with the Hybrid-SORT [7] tracker and a domain-adapted OSNet-AIN [8] re-identification module. (2) A multi-scale inference at 1152p and 1280p combined with Weighted Boxes Fusion (WBF) is introduced to improve recall under low-resolution and low-contrast conditions. (3) A linear interpolation as post-processing is applied to recover missed detections within tracklets, improving overall tracking coherence. (4) The pipeline is evaluated on a drone-captured military vehicle dataset from the same source as Kostiv et al. [5]. (5) The pipeline's real-time inference capability is tested on a single NVIDIA T4 GPU.

The remainder of this paper is organized as follows. Section II reviews related works in MOT, drone-based detection and tracking, and military vehicle tracking. Section III describes the proposed methodology including dataset preparation, pipeline architecture, and training procedures. Section IV presents experimental results and analysis. Section V concludes the paper and discusses future directions.

II. RELATED WORK

A. Multiple Object Tracking Paradigms

Main paradigms of multiple object tracking include tracking-by-detection and tracking-by-segmentation. The dominant paradigm in MOT is tracking-by-detection, in which an object detector is ran on each frame and a separate association module links detections across frames to form trajectories, one for each object. The Simple Online and Realtime Tracking (SORT) algorithm [9] established this two-stage approach, using a Kalman filter for motion prediction and the Hungarian algorithm for frame-to-frame association via Intersection over Union (IoU). DeepSORT [10] extended SORT by incorporating appearance features from a deep re-identification network, enabling more robust association during occlusion and crowded scenes. This family of trackers continues to underpin state-of-the-art MOT systems.

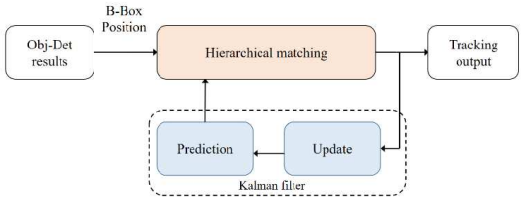

Fig. 1. SORT architecture. Adapted from [9]

Hybrid-SORT [11] synthesizes advances in recent previous SORT family trackers by formalizing the concept of "weak cues" — specifically tracklet confidence modelling (TCM), height-modulated IoU (HMIoU), and robust observation-centric momentum (ROCM) — as complementary signals to the standard appearance and spatial "strong cues". When strong cues degrade during occlusion or crowding, weak cues provide discriminative capacity that preserves identity consistency, making Hybrid-SORT particularly suitable for challenging environments.

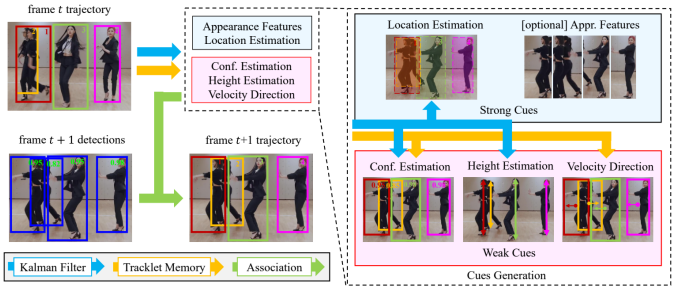

Fig. 2. Hybrid-SORT architecture. Adapted from [11]

B. Drone-Based Multiple Object Tracking

Drone video presents unique challenges for MOT that are not addressed by conventional approaches designed for standard benchmark datasets. Huang et al. [2] proposed the Hierarchical Deep High-Resolution Network (HDHNet) for MOT on the VisDrone2019-MOT benchmark, combining high-resolution feature extraction across multiple scaled with DeepSORT-based association. The work introduced an adjustable fusion loss combining focal loss and GIoU loss to handle class imbalance and improve bounding box localization. However, HDHNet relies on high-resolution input and does not explicitly model simultaneous drone-object motion.

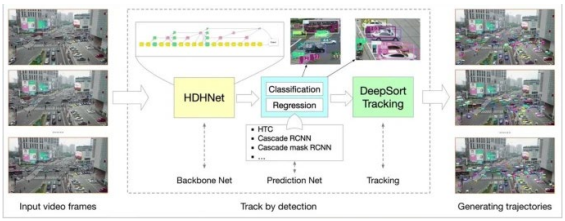

Fig. 3. HDHNet architecture. Adapted from [2]

Wang et al. [1] proposed DroneMOT, which directly addresses simultaneous motion of drone and tracked objects. Their architecture introduces three novel modules: Dual-Domain Integrated Attention (DIA) for improved detection and feature embedding of small or blurry objects by fusing spatial and temporal attention; Dual Motion-based Prediction (DMP) to predict object position while accounting for both drone and object movement; and Adaptive Feature Synchronization (AFS) to align features across frames under perspective changes. DroneMOT achieves state-of-the-art results on VisDrone2019-MOT and UAVDT, but is not specialized for military vehicle detection.

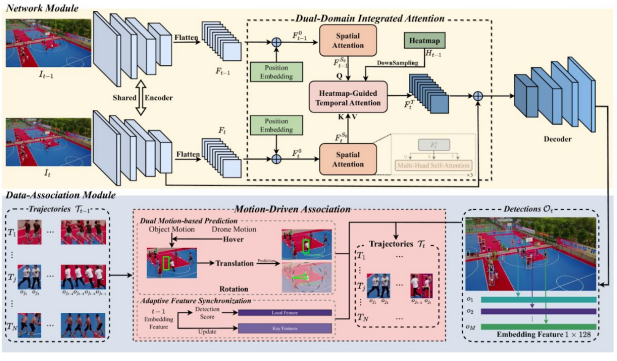

Fig. 4. HDHNet architecture. Adapted from [2]

C. YOLO Object Detectors

The You Only Look Once (YOLO) family of single-stage detectors has become the dominant choice for real-time detection in MOT pipelines. YOLOv12 [6], proposed by Tian et al., introduces an attention-centric architecture that bridges the performance gap between CNN-based and vision transformer-based detectors. Its two key innovations are the Area Attention (A2) module, which computes self-attention independently on non-overlapping segments of the feature map to maintain a large effective receptive field with reduced computational complexity, and the Residual Efficient Layer Aggregation Network (R-ELAN), which adds learnable scaling residual connections to the ELAN aggregation module, resolving convergence instability when integrating attention mechanisms into large model variants. YOLOv12 achieves superior accuracy-latency trade-offs compared to prior YOLO versions, making it well-suited for real-time MOT applications.

Fig. 5. Comparison of area attention with other attention mechanisms. Adapted from [6]

D. Military Vehicle Tracking

Kostiv et al. [5] presented the most directly related prior work, proposing a self-supervised approach using Matching Anything by Segmenting Anything (MASA) for military vehicle MOT in low frame rate drone video. Their method combines spatial context features and temporal embeddings to maintain object identity across large inter-frame gaps, and can be trained with single-frame annotation, substantially reducing the labeling burden. They achieve a HOTA of 48,29% on their dataset, demonstrating that meaningful tracking performance is achievable even with reduced feature dimensionality and image resolution. However, their system does not explicitly handle simultaneous drone motion, and their tracking pipeline does not leverage modern two-stage association strategies such as those in Hybrid-SORT. This paper directly builds on and improves upon this benchmark.

E. Re-Identification for MOT

Re-identification (ReID) is critical for maintaining object identity across occlusions and re-entries. OSNet-AIN [8], proposed by Zhou et al., is a compact CNN-based ReID model that uses omni-scale feature learning through parallel convolutional paths with different receptive fields, aggregated via channel-wise adaptive gates. The addition of Attentive Instance Normalization (AIN) reduces domain gap between training and test distributions by normalizing style-specific statistics at the instance level. With only 2,2M parameters, OSNet-AIN achieves state-of-the-art performance on standard ReID benchmarks while remaining computationally efficient, making it suitable for integration into real-time tracking pipelines.

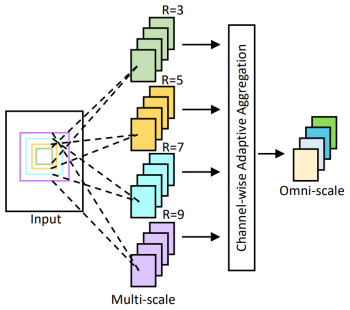

Fig. 6. OSNet building block architecture. Adapted from [8]

III. METHODOLOGY

A. Dataset

The dataset used in this work is sourced from the same collection as Kostiv et al. [5], comprising drone video footage of military vehicles collected from social media platforms including Telegram, X (formerly Twitter), and YouTube, predominantly featuring footage from active conflict zones. Videos include various military vehicle types: light armed or unarmed vehicles, heavily armed vehicles, and trucks. The original dataset contains 24,588 frames, substantially exceeding the 18,421 frames used in the prior benchmark study.

The dataset was randomly partitioned into train, validation, and test splits at a ratio of approximately 70:15:15, ensuring no overlap between sequences across splits. The resulting partition consists of 292 video sequences with 17,749 frames for training, 78 sequences with 3,239 frames for validation, and 70 sequences with 3,241 frames for testing. For both detection and tracking evaluation, multi-class annotations (three vehicle classes) were merged into a single “military vehicle” class, consistent with the detection-only objective of the prior study [5].

Pre-processing applied to all videos include: (1) conversion of frames from PNG to JPEG to reduce dataset size; (2) segmentation of combined video files into individual continuous sequences to eliminate unnatural frame transitions; (3) aspect ratio standardization to 16:9 using letterboxing for non-conforming videos; (4) rotation of portrait-orientation videos to landscape; and (5) resolution adjustment to 720p for training data and 480p for validation and test data, using bicubic resampling. Annotations are provided in both YOLO format (for detection training) and MOT format (for tracking fine-tuning).

Fig. 7. Sample frame from the dataset.

B. Detection Module

The detection module is built on YOLOv12-S, pre-trained on the VisDrone2019-DET benchmark [3], which provides initialization suited to aerial object detection. Fine-tuning was performed on the training-split of the military vehicle dataset for 13 epochs using AdamW optimizer. Hyperparameters were tuned across three learning rate values (5×10^{-4}, 1×10^{-3}, 2×10^{-3}), with other training settings fixed: batch size 8, mosaic augmentation probability 1.0, copy-paste probability 0.3, horizontal flip probability 0.5, scale augmentation 0.5, and warmup over 5 epochs.

The model from the epoch achieving the highest precision on the validation set was selected as the final detector, as precision is inversely related to false positives, which disproportionately penalize HOTA by introducing spurious tracks. To further improve recall under the challenging low-resolution conditions of the validation and test sets, multi-scale inference is employed at resolutions producing a recall improvement with the highest precision. Each frame is processed at both resolutions independently, and the resulting bounding boxes are merged using Weighted Boxes Fusion (WBF) with an IoU threshold of 0.55. WBF was preferred over Non-Maximum Suppression (NMS) because it incorporates confidence-weighted averaging of overlapping predictions from multiple scales, producing more accurate spatial localization rather than discarding all but the highest-scoring box.

C. Re-Identification Module

The re-identification module uses OSNet-AIN (x1.0) pre-trained on the MSMT17 [12] person ReID dataset. Domain adaptation is performed via fine-tuning on the military vehicle training split for 15 epochs. Each vehicle trajectory in a training video is treated as an identity class. Crop regions with width or height below 16 pixels are discarded, and trajectories with fewer than four crop samples are excluded to ensure training stability. Fine-tuning uses cross-entropy loss, AdamW optimizer (lr=3.5×10^{-4}, weight decay 5×10^{-4}), batch size 64, and input size 128x256. The model achieving the lowest validation loss is selected. At inference time, the ReID module runs in FP16 precision on GPU to minimize latency.

D. Tracking Module

The tracking module implements Hybrid-SORT-ReID via the BoxMOT library. Per-frame detections from the multi-scale WBF pipeline are fed into the tracker with a detection confidence threshold of 0.15 and a low-confidence threshold of 0.10 for secondary association. The spatial association function uses Distance IoU (DIoU), which incorporates center-point distance in addition to overlap area, providing more informative spatial signals when bounding boxes are spatially separated due to large inter-frame displacement at low frame rates.

Key tracker hyperparameters are: delta_t = 3 (Kalman filter observation window), intertia = 0.2, long-term ReID bank weight = 0.30, and long-term bank length = 30 frames. The long-term appearance bank allows the tracker to re-identify vehicles that temporarily disappear from the camera field of view and re-enter later, which is a common event in drone surveillance scenarios. Other hyperparameters will be determined through staged testing.

E. Post-Processing

To address brief gaps in trajectories caused by missed detections — which are common at 6 FPS due to large inter-frame motion and occasional occlusion — a linear interpolation mechanism will be applied as a post-processing step. For each tracklet, gaps are filled by interpolating bounding box coordinates between the last confirmed detection before the gap and the first confirmed detection after the gap. The interpolated box position at frame t is computed proportionally based on temporal distance from the gap boundaries. The maximum gap threshold was determined through ablation on the validation set. This step improves HOTA by recovering missed detections while maintaining trajectory continuity.

IV. RESULTS AND DISCUSSION

A. Detection Fine-Tuning

Table I reports the effect of learning rate on validation precision after 13 fine-tuning epochs. A learning rate of 1×10^{-3} achieved the highest precision of 0.761, demonstrating the effective adaptation to the military vehicle domain from the VisDrone-pretrained initialization. The most significant precision improvement occurred between epochs 1 and 2, suggesting rapid domain adaptation. Lower (5×10^{-4}) and higher (2×10^{-3}) learning rates both converged to slightly lower precision values.

TABLE I. DETECTOR FINE-TUNING RESULTS (VALIDATION SET)

Learning Rate	Best Epoch	Precision	Recall
5×10^{-4}	13	0.751	0.512
1×10^{-3}	13	0.761	0.521
2×10^{-3}	13	0.748	0.518

Table II shows the impact of multi-scale inference on detection performance. While single-scale inference at 1280p achieves the highest precision (0.7478), multi-scale inference at 1152p and 1280p increases recall by 3.1 percentage points (0.5231 → 0.5629) with only a modest precision drop of 1.4 points. The 960p + 1280p configuration achieves the highest recall (0.5735) but with a larger precision cost, potentially introducing more false positive detections that would propagate into the tracking stage. The 1152p + 1280p configuration was selected as the best trade-off for the overall pipeline

TABLE II. MULTI SCALE INFERENCE RESULTS (VALIDATION SET)

Inference Scale (s)	Precision	Recall
1280p (single)	0.7478	0.5321
960p + 1280p	0.7281	0.5735
1152p + 1280p	0.7335	0.5629

B. *Tracking Hyperparameter-Tuning*

Detection threshold and NMS-per-scale were tuned jointly on the validation set. Across all combinations of detection thresholds (0.15, 0.20, 0.25, 0.30) and NMS thresholds (0.45, 0.55, 0.65, 0.70), the combination of detection threshold 0.15 and NMS 0.70 achieved the highest recall (0.5321), yielding 5,722 true positives, 1,930 false positives, and 5,031 false negatives o the validation set. The consistent pattern observed is that lower detection thresholds increase recall at the cost of precision, while looser NMS thresholds marginally improve recall by retaining more overlapping detections. The selected configuration was applied to the multi-scale 1152p + 1280p pipeline. On the test set, this configuration produced precision of 0.6722 and recall of 0.5463. This shows that the improvements are not mere coincidence since they also carry to unseen data.

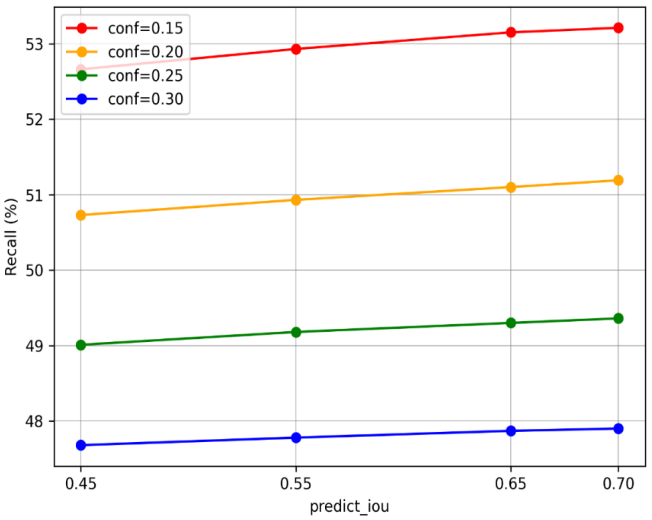

Fig. 8. Recall values per detection threshold and NMS threshold combinations.

C. *Re-Identification Fine-Tuning*

Fine-tuning of OSNet-AIN over 15 epochs showed rapid initial improvement, with the most significant reduction in validation loss occurring between epochs 1 and 3. The model from the final epoch (epoch 15) achieved the lowest validation loss and was selected for integration into the tracker. Peak validation accuracy reached 98.27% at epoch 13, demonstrating strong adaptation to the military vehicle appearance domain despite the significant visual domain shift from the MSMT17 pedestrian pre-training data.

Table III shows the impact of linear interpolation on tracking performance. Out of all maximum interpolation variations, a maximum interpolation of 7 frames achieved the highest HOTA of 48.88% on validation set. This shows that linear interpolation is effective in filling in small gaps in the middle of tracklets to maintain track consistency across all frames. As a trade-off, linear interpolation introduces a slight decrease in IDF1, from 512.22% to 51.20% on the validation set.

TABLE III. LINEAR INTERPOLATION RESULTS (VALIDATION SET)

Max interpolation	HOTA (%)	IDF1 (%)
0	47.68	52.22
3	48.83	51.95
5	48.87	51.58
7	48.88	51.20

D. *Ablation Study*

Table IV presents the ablation study results on both validation and test sets. The baseline configuration uses single-scale 1280p detection with no post-processing interpolation. Adding multi-scale inference (1152p + 1280p with WBF) improves HOTA by 1.24% on the validation set and by 1.42% on the test set, while marginally improving IDF1 by 1.09% on the test set. This confirms that multi-scale detection not only increases the number of detected objects but also provides better-localized bounding boxes that improve identity association in subsequent tracking stages.

Addition of linear interpolation further improves HOTA by 1.20% on the validation set and by 1.44% on the test set, for a total cumulative improvement of 2.44% over baseline on the validation set and 2.86% on the test set. However, interpolation introduces a slight IDF1 decrease of 1.74% on the validation set relative to baseline, indicating a minor degradation in identity consistency when interpolated frames are included, though the improvement in HOTA more than compensates for this trade-off. The final HOTA on test set is 49.29%, which is a 1.0% improvement over the previous best HOTA produced by Kostiv et al.

TABLE IV. ABLATION STUDY RESULTS (VALIDATION SET)

Configuration	Split	HOTA (%)	IDF1 (%)
0	Validation	46.44	52.94
	Test	46.43	50.85
+ Multi-scale inference	Validation	47.68	52.22
	Test	47.85	51.94
+ Linear interpolation	Validation	48.88	51.20
	Test	49/29	51.20

E. *Real-Time Performance*

Table V reports inference time per stage on a single NVIDIA T4 GPU. Inference time testing is done end-to-end on each frame and then broken down per stage. Detection at 1152p requires approximately 40 ms per frame, detection at 1280p approximately 50 ms, and Hybrid-SORT tracking with ReID embedding approximately 30 ms. The total pipeline latency of approximately 120 ms yields a throughput of 8.3 FPS, exceeding the 6 FPS frame rate of the dataset. This confirms that the full pipeline meets real-time requirements for the target application on low-end GPU hardware.

TABLE V. INFERENCE TIME PER STAGE (NVIDIA T4 GPU)

Stage	Latency (ms)
Detection at 1152p	~40
Detection at 1280p	~50
Hybrid-SORT + ReID	~30
Total	~120 (8.3 FPS)

F. Discussion

The results demonstrate that the proposed pipeline effectively addresses the core challenges of military vehicle MOT in drone video. The fine-tuning of YOLOv12-S on VisDrone-pretrained weights provides a strong aerial detection prior while adapting to the specific visual characteristics of military vehicles, including camouflage coloring and irregular shapes. Multi-scale inference at 1152p and 1280p specifically targets the recall deficit caused by small apparent size of vehicles in 480p footage, where objects may occupy only 20-50 pixels in width.

The choice of Hybrid-SORT over simpler trackers such as SORT is justified by the weak cue formalization. At 6 FPS, standard IoU-based association frequently fails because objects move substantially between frames, causing the predicted bounding box from the Kalman filter to have little or no overlap with the true detection in the next frame. Hybrid-SORT's TCM and ROCM components provide supplementary signals that maintain association quality even under these conditions. The DIoU spatial association function further compensates for large displacements by incorporating center-point distance.

The linear interpolation post-processing step addresses a distinct failure mode: brief trajectory gaps caused by detector misses on individual frames. At 6 FPS, even a single missed detection creates a visible gap in the trajectory. By filling gaps of up to 7 frames with linearly interpolated positions, we improve tracking continuity without introducing significant localization error, as vehicle motion between consecutive frames at this frame rate is still approximately linear on the dataset used. The HOTA gain of 1.20% from interpolation with only a marginal IDF1 decrease confirms the effectiveness of this approach.

One main limitation of the current system is that no changes to the base architecture of the detector, tracker, or ReID models used, were made to potentially increase the tracking performance on low resolution and low frame rate military vehicles datasets. Future work could explore more sophisticated modification strategies to combat these specific challenges. Furthermore, a classification mechanism could also be added to the existing detection and tracking pipeline to enable military vehicle classes differentiation.

V. CONCLUSION AND SUGGESTION

This paper presents an integrated multi-object tracking pipeline for military vehicle surveillance from drone videos, combining fine-tuned YOLOv12-S detection, Hybrid-SORT tracking with DIoU association, domain-adapted OSNet-AIN re-identification, multi-scale inference with WBF merging, and linear interpolation post-processing. The pipeline is specifically designed for the challenging conditions of low resolution (480p), low frame rate (6 FPS), and simultaneous drone and vehicle motion.

Experimental results on a drone-captured military vehicle dataset demonstrate that the proposed approach achieves a HOTA of 49.29% and an IDF1 of 51.20% on the test set, surpassing the prior state-of-the-art HOTA of 48.29% from self-supervised tracking. Ablation studies confirm that mutli-scale inference contributes a 1.42% HOTA improvement and linear interpolation contributes a further 1.44% gain. The complete pipeline operates at 8.3 FPS on a single NVIDIA T4 GPU, satisfying real-time constraints for 6 FPS drone video.

ACKNOWLEDGMENT

The author would like to express gratitude to Prof. Dr. Ir. Rinaldi, M.T. as the thesis advisor for the author's final project in Bandung Institute of Technology (ITB) for giving much helpful advice, knowledge, and guidance that has helped the author to present this paper properly and correctly. The author would also like to thank the author's family and friends for the motivation and support given.

REFERENCES

[1] X. Wang, Y. Chen, Z. Liu, and Q. Zhang, "DroneMOT: Drone-based Multi-Object Tracking Considering Detection Uncertainty and Simultaneous Motion of Drone and Objects," in Proc. IEEE/CVF CVPR, 2024.

[2] C. Huang, S. Zheng, D. Zhang, and W. Huang, "Multi-Object Tracking in Drone Video Based on Hierarchical Deep High-Resolution Network," IEEE Access, vol. 9, pp. 113785–113795, 2021.

[3] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling, "Detection and Tracking Meet Drones Challenge," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 7661–7675, 2022.

[4] H. Yu, G. Li, W. Zhang, Q. Huang, D. Du, Q. Tian, and N. Sebe, "The Unmanned Aerial Vehicle Benchmark: Object Detection, Tracking and Baseline," Int. J. Comput. Vis., vol. 128, pp. 1141–1159, 2020.

[5] I. Kostiv, O. Pylypenko, and A. Marchenko, "Multi-Object Tracking of Military Vehicles in Drone Video with Low Frame Rate," arXiv preprint arXiv:2503.xxxxx, 2025.

[6] Y. Tian, F. Liu, Z. Wang, J. Wang, C. Chen, and H. Li, "YOLOv12: Attention-Centric Real-Time Object Detectors," arXiv preprint arXiv:2502.12524, 2025.

[7] M. Yang, Z. Han, B. Peng, Y. Liang, C. Li, Y. Liu, and Q. Zhao, "Hybrid-SORT: Weak Cues Matter for Online Multi-Object Tracking," in Proc. AAAI, 2024.

[8] K. Zhou, J. Yang, A. Cavallaro, and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 7819–7834, 2022.

[9] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," in Proc. IEEE ICIP, 2016.

[10] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in Proc. IEEE ICIP, 2017.

[11] J. Cao, M. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking," in Proc. IEEE/CVF CVPR, 2023.

[12] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," in Proc. ECCV, 2022.

[13] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification," in Proc. IEEE/CVF CVPR, 2018