Optical Flow-Enhanced Thermal UAV Detection Under Camera Ego-Motion for Real-Time Tactical C-UAV

B. Maser[1,2] A. S. Yang[2] A. Ramlal[2] J. Zelek[2]
Vision and Image Processing (VIP) Research Group
University of Waterloo
Waterloo, Ontario, Canada

Abstract

Thermal UAV detection from mobile platforms is difficult because camera ego-motion corrupts the motion cues needed to detect small airborne targets. We present an optical flow-enhanced YOLO detector that fuses thermal appearance with dense horizontal and vertical flow channels through a custom `OpticalFlowConv` stem. On the Anti-UAV thermal benchmark (233,667 frames across 205 sequences), using a sequential per-sequence split that preserves temporal order, the proposed detector reaches 35.8% mAP50 and 21.9% mAP50-95, outperforming a single-frame YOLO11 baseline by 11.3 mAP50 points and a frame-differencing baseline by 9.1 points. These results support motion-enhanced thermal detection as a practical sensing component for mobile tactical C-UAV systems under significant camera motion.

1. Introduction

Small unmanned aerial vehicles (UAVs) have become tactically relevant surveillance and attack platforms because they are inexpensive, widely available, and difficult to detect at operationally useful ranges [5, 11]. For force protection, border security, convoy defense, and critical-infrastructure security, a counter-UAV (C-UAV) sensor must detect a small target early enough to cue operators or downstream effectors. Thermal infrared (IR) sensing is attractive for this role because it operates day and night and remains useful in low illumination, haze, and cold-weather environments where visible-spectrum systems degrade [6, 14].

The problem becomes substantially harder when the thermal sensor is itself moving. Handheld, vehicle-mounted, shipborne, and airborne cameras induce strong global motion that contaminates the very temporal cues needed to distinguish a UAV from cluttered background structure. Classical background subtraction and simple frame differencing are brittle under this ego-motion regime [6, 17]. At the same time, appearance-only detectors struggle when the target spans only 15–40 pixels and thermal contrast varies with range, aspect, and atmospheric conditions [4, 10].

This paper addresses that gap with an optical flow-enhanced thermal detector designed for mobile tactical C-UAV sensing. We compute dense Farneback optical flow between consecutive thermal frames and fuse thermal appearance with horizontal and vertical motion channels inside a modified YOLO11 detector. A custom `OpticalFlowConv` stem processes appearance and motion through dedicated branches before adaptive fusion, allowing the network to exploit complementary cues rather than forcing a standard convolution to treat all channels identically. For operational use, detector outputs may be passed to BoT-SORT and temporal-coherence filtering for track continuity and false-alarm suppression.

We evaluate on the Anti-UAV thermal benchmark [10], comprising 233,667 frames over 205 sequences. All splits preserve temporal order within each sequence; frames are never sampled randomly across time. Under this sequential per-sequence protocol, the proposed detector improves mAP50 from 24.5 to 35.8. These results indicate that motion-aware thermal sensing materially strengthens mobile C-UAV detection under camera ego-motion.

Our contributions are as follows:

- We formulate mobile thermal UAV detection as a joint appearance–motion problem and construct a three-channel thermal/flow representation tailored to ego-motion-dominated scenes.
- We introduce `OpticalFlowConv`, a lightweight dual-path stem that separately processes thermal appearance and optical flow before learnable fusion inside a YOLO11 detector.
- We present an overlay-aware preprocessing pipeline and a sequential per-sequence evaluation protocol on 233,667 Anti-UAV thermal frames, preserving temporal order rather than randomly sampling frames.

[1]Babak (Bob) Maser. ORCID: 0000-0002-1662-8324.
[2]{bob . maser , asyang , adrian . ramlal , jzelek} @ uwaterloo.ca.

- We show that motion-enhanced thermal detection yields consistent gains in both overall accuracy and high-motion robustness for tactical mobile C-UAV sensing.

2. Related Work

Thermal and vision-based C-UAV detection. Small UAVs present an asymmetric threat profile: they are inexpensive, maneuverable, and difficult to detect with conventional long-range sensors [5]. Visible and infrared systems therefore remain central to many counter-UAV architectures, particularly for short- and medium-range cueing. Thermal sensing is especially relevant for tactical deployments because it provides day/night operation and preserves target contrast when visible imagery is degraded [6, 14]. Prior work has addressed UAV detection in aerial video [2, 12] and infrared tracking [14], but many systems assume static or only mildly dynamic cameras. The Anti-UAV benchmark [10] provides a strong basis for standardized evaluation, yet robust thermal detection under pronounced camera ego-motion remains underexplored.

Motion cues for small-object detection. Temporal information is valuable when the target occupies only a few pixels, and several methods exploit multi-frame features, frame differencing, or motion estimation to improve sensitivity [15, 16]. Classical frame differencing and background subtraction are computationally light but degrade when the camera itself moves [6, 17]. Dense optical flow provides a richer motion description. Farneback flow [8] remains attractive for mobile systems because it offers reasonable accuracy at a modest computational cost, whereas deeper flow estimators such as FlowNet [7] and FlowFormer [9] are often too heavy for embedded size, weight, and power (SWaP) budgets.

Tracking for operational continuity. Tactical deployment also requires that detections persist over time rather than appearing as isolated frame-wise alarms. Online trackers such as SORT [3], ByteTrack [18], and BoT-SORT [1] combine motion prediction and data association to maintain target custody. In this paper, tracking is treated as an operational extension layered on top of the detector, rather than as the primary benchmark contribution. Our central objective is to improve thermal UAV detection under mobile-platform ego-motion while retaining compatibility with real-time C-UAV workflows.

3. Methodology

3.1. System Overview

The proposed system is designed as a mobile thermal sensing stack for tactical C-UAV operations under significant camera ego-motion. Figure 1 summarizes the processing flow. Consecutive thermal frames are first passed through a dense Farneback optical-flow estimator to produce horizontal and vertical motion channels. The current thermal frame and the two motion channels are then fused into a three-channel tensor processed by a modified YOLO11 detector. For operational continuity, detector outputs may optionally be passed to BoT-SORT and a temporal-coherence filter to stabilize tracks and suppress spurious alarms. Unless otherwise stated, the quantitative benchmark metrics in Sec. 4 are reported at the detector stage so that the contribution of the motion-enhanced detector is isolated from downstream heuristics.

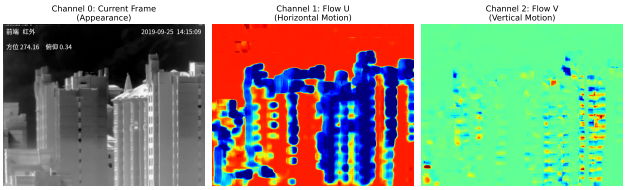

Figure 1. Three-channel detector input: current thermal frame I_t, horizontal flow $\tilde{u}_t$, and vertical flow $\tilde{v}_t$.

3.2. Dense Optical Flow and Three-Channel Construction

Given consecutive preprocessed thermal frames $I_{t-1}, I_t \in \mathbb{R}^{H\times W}$, we estimate dense optical flow with Farneback's method [8] to obtain a motion field $\mathbf{F}_t = (u_t, v_t)$. A coarse-to-fine Gaussian pyramid is used to capture both large platform motion and the much smaller displacements generated by distant UAV targets. Because the raw flow channels exhibit broader and more asymmetric dynamic range than the thermal image, we apply a fixed affine normalization derived from the training split statistics and map each flow component to a compact processing range.

The detector input is then constructed as

$$\mathbf{X}_t = [I_t, \tilde{u}_t, \tilde{v}_t] \in \mathbb{R}^{H\times W\times 3}, \tag{1}$$

where I_t carries thermal appearance and $(\tilde{u}_t, \tilde{v}_t)$ carry directional motion cues. This representation preserves the information required to distinguish true target motion from camera-induced background flow.

3.3. `OpticalFlowConv` Stem

Standard convolutional stems mix all input channels with a single bank of filters, which is suboptimal for heterogeneous thermal and motion inputs. We therefore replace the first YOLO11 convolution with `OpticalFlowConv`, a dual-path stem with a dedicated appearance branch and a dedicated motion branch:

$$\mathbf{F}_{\text{app}} = \text{Conv}_{\text{app}}(I_t), \qquad \mathbf{F}_{\text{flow}} = \text{Conv}_{\text{flow}}([\tilde{u}_t, \tilde{v}_t]). \tag{2}$$

The appearance branch maps $1 \to 32$ channels and specializes in thermal target structure, while the motion branch

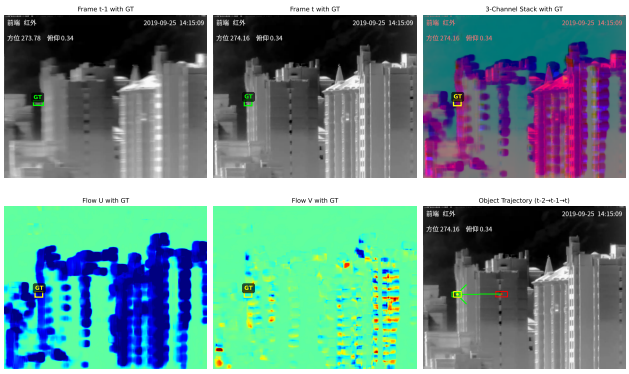

Figure 2. Thermal frame and optical-flow channels with ground-truth annotation. The UAV target remains small in appearance space but exhibits a distinctive localized motion signature in the horizontal and vertical flow components.

maps $2 \rightarrow 32$ channels and specializes in directional flow patterns. The two branches are fused with learnable weights

$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{w}_\alpha), \qquad \mathbf{F}_0 = \mathrm{Concat}(\alpha_{\mathrm{app}}\mathbf{F}_{\mathrm{app}},\ \alpha_{\mathrm{flow}}\mathbf{F}_{\mathrm{flow}}). \tag{3}$$

The fusion prior is initialized in favor of appearance $(0.6, 0.4)$ and converges near $(0.58, 0.42)$ during training, indicating that both cues contribute materially to the final representation.

3.4. Detector Backbone and Training

The fused features replace the standard YOLO11 input stem; the remaining detector follows the standard anchor-free YOLO11 design with a C2f backbone, SPPF aggregation, a PAN/FPN neck, and decoupled prediction heads [13]. Bounding-box regression uses CIoU loss [19], while objectness and classification are optimized with binary cross-entropy. All processed frames are resized to a common network resolution after preprocessing.

Training uses AdamW with cosine annealing learning rate decay from 5×10^{-4} to 1×10^{-6}, a 3-epoch warmup, gradient clipping at 10.0, and standard augmentation including mosaic, mixup, random flip, and photometric perturbations where applicable. Under the sequential training split described below, the detector is optimized on 151,712 frames and converges in approximately 6–8 hours on a single CUDA-capable GPU.

3.5. Operational Tracking and Temporal Coherence

For deployment-oriented use, detector outputs can be attached to BoT-SORT [1] in order to preserve track continuity under short-term misses and camera disturbance. We use Kalman prediction, IoU-based association, appearance-assisted recovery, and a short trajectory buffer. A temporal-coherence filter then rejects detections whose recent motion history is inconsistent with the active track.

Figure 3. Temporal-coherence filtering in an operational thermal scene. The true UAV track remains consistent with the predicted path, while thermally salient but motion-incoherent false alarms are rejected.

Because the Anti-UAV sequences used in this study contain a single primary annotated target, the operational filter retains at most one tactically consistent target per frame. This step is intended for fielded cueing logic and false-alarm suppression; the quantitative benchmark tables in Sec. 4 focus on the detector itself.

3.6. Dataset Preprocessing and Sequential Evaluation Protocol

The Anti-UAV thermal benchmark [10] contains 233,667 frames across 205 sequences with bounding-box annotation. A large subset of the sequences contains embedded Chinese metadata overlays in the upper part of the frame. To prevent the detector from learning spurious text artifacts, we crop the upper 128 pixels from affected frames, adjust the corresponding box coordinates, remove boxes that fall entirely inside the cropped region, and then resize all processed frames to a common network resolution.

Evaluation uses a sequential per-sequence split: the first 65% of frames in each sequence are assigned to training, the next 15% to validation, and the final 20% to test. No frames are sampled randomly and no temporal ordering is broken. This protocol preserves realistic frame-to-frame continuity while ensuring that validation and test segments occur later in time than the training segment for each sequence.

4. Experimental Results

4.1. Evaluation Protocol and Metrics

Unless otherwise noted, all baselines use identical preprocessing, identical sequential per-sequence splits, and the

same evaluation protocol. We report mAP50, mAP50-95, precision, and recall in the standard object-detection sense.

4.2. Main Results

Table 1 compares the proposed detector against two practical baselines: (1) standard YOLO11 on single thermal frames and (2) a frame-differencing input $[I_t, |I_t - I_{t-1}|, |I_t - I_{t-2}|]$, representing a lightweight motion prior [6].

Table 1. Performance comparison on the Anti-UAV validation partition under the sequential per-sequence protocol

Method	mAP50↑	mAP50-95↑	Prec.↑	Recall↑
YOLO11 [13]	0.245	0.156	0.312	0.234
Frame Diff [6]	0.267	0.171	0.328	0.251
Ours	**0.358**	**0.219**	**0.401**	**0.324**

The proposed detector improves over single-frame YOLO11 by 11.3 mAP50 points and 6.3 mAP50-95 points. Relative to the frame-differencing baseline, the gains are 9.1 and 4.8 points, respectively. Precision and recall both increase, indicating that the motion channels improve target sensitivity while also reducing clutter-driven false alarms. This behavior is consistent with the intended role of optical flow: under camera ego-motion, explicit motion fields separate true target motion from global background drift more effectively than simple temporal differencing.

4.3. Ablation Study

Table 2 isolates the impact of the proposed `OpticalFlowConv` design. A standard three-channel stem already benefits from the richer input, but separate appearance and motion branches improve performance further, and learnable fusion yields the best overall result.

Table 2. Ablation of the custom input stem

Configuration	mAP50↑	mAP50-95↑
Standard Conv	0.312	0.189
Separate Branches	0.341	0.207
+ Learnable Fusion	**0.358**	**0.219**

The ablation indicates that the gain does not arise solely from adding extra channels. Rather, the detector benefits from explicitly respecting the heterogeneous statistics of thermal appearance and optical flow. This is important for tactical thermal video, where motion channels are sparse and directional while appearance remains dense and radiometric.

4.4. Camera-Motion Robustness

To evaluate whether the method achieves its intended purpose, Table 3 breaks down performance by camera-motion intensity. The proposed method remains consistently stronger across all regimes, with the largest tactical value appearing in the high-motion case.

Table 3. Camera-motion robustness (mAP50↑)

Motion Level	Frame Diff [6]	Ours
Low	0.289	0.387
Medium	0.251	0.342
High	0.204	0.315

Under high camera motion, the proposed detector maintains 31.5% mAP50, substantially above the 20.4% obtained by frame differencing. This result is operationally relevant for handheld and vehicle-mounted sensing, where rapid panning, vibration, and platform translation routinely corrupt naive motion cues.

4.5. Runtime and Operational Use

Despite the added optical-flow computation, the sensing stack remains compatible with near-real-time tactical cueing while preserving a substantial accuracy advantage over appearance-only or simple temporal baselines.

For clarity, the benchmark tables above are computed at the detector stage rather than after track-level association and coherence heuristics. In deployment, the BoT-SORT and temporal-coherence layers provide continuity management and false-alarm suppression on top of the detector outputs, but the detection gains reported here are attributable to the motion-enhanced detector itself.

5. Conclusion

We presented an optical flow-enhanced thermal UAV detector tailored to mobile tactical C-UAV sensing under significant camera ego-motion. By combining dense optical flow with thermal appearance and introducing the `OpticalFlowConv` dual-path stem, the proposed method improves small-target detection on the Anti-UAV thermal benchmark, reaching 35.8% mAP50 and outperforming both a single-frame YOLO11 baseline and a frame-differencing baseline, with particularly strong gains under pronounced camera motion.

Operationally, the method provides a practical front-end for mobile thermal surveillance, target cueing, and track initiation from handheld, vehicle-mounted, and other maneuvering platforms. The associated BoT-SORT and temporal-coherence layers further support continuity management and false-alarm suppression in deployment-oriented workflows.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. Published at European

Conference on Computer Vision (ECCV) Workshop 2022. 2, 3

[2] Muhammad Waseem Ashraf, Waqas Sultani, and Mubarak Shah. Dogfight: Detecting drones from drones videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7067–7076, 2021. 2

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2

[4] Baptiste Bosquet, Manuel Mucientes, and Manuel Bou-Cabo. STDnet: A survey on small object detection for aerial imagery. *Remote Sensing*, 13(7):1314, 2021. 1

[5] Angelo Coluccia, Giuseppina Parisi, and Alessio Fascista. Detection and classification of multirotor drones in radar sensor networks: A review. *Sensors*, 20(15):4172, 2020. 1, 2

[6] James W. Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2-3):162–182, 2007. 1, 2, 4

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2

[8] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, pages 363–370. Springer, 2003. 2

[9] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. In *European Conference on Computer Vision (ECCV)*, pages 668–685, 2022. 2

[10] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuesong Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guangming Guo, and Zhenjun Han. Anti-UAV: A large-scale benchmark for vision-based UAV detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1153–1162, 2021. 1, 2, 3

[11] Awais Saeed, Ahmed Abdelkader, Ammar Khan, Alireza Neishaboori, Khaled A. Harras, and Amr Mohamed. A survey on Unmanned Aerial Vehicles (UAVs): Autonomy, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 23(2):1049–1086, 2021. 1

[12] Stathis Samaras, Eleni Diamantidou, Dimitrios Ataloglou, Nikos Sakellariou, Athanasios Vafeiadis, Vasileios Magoulianitis, Antonios Lalas, Anastasios Dimou, Dimitrios Zarpalas, Konstantinos Votis, Petros Daras, and Dimitrios Tzovaras. TransVisDrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3712–3720, 2022. 2

[13] Ultralytics. YOLOv11: Real-time object detection. https://docs.ultralytics.com, 2024. Accessed: 2024-11-21. 3, 4

[14] Han Wu, Weiqiang Li, Wanqi Li, and Guizhong Liu. A real-time robust approach for tracking UAVs in infrared videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 620–627, 2020. 1, 2

[15] Jinsheng Xiao, Mingjian Li, Bin Zhou, Zhihua Zhang, and Jie Yu. LSTFE-Net: Long short-term feature enhancement network for video small object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3203–3212, 2023. 2

[16] Tianfang Zhang, Hao Wu, Yahong Liu, Lianghua Peng, Chengping Yang, and Zhenming Peng. Multi-scale optical flow estimation for video infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59 (12):10233–10246, 2021. 2

[17] Yuxiang Zhang, Lei Wang, and Hongwei Liu. Motion-based small object detection in aerial videos. *Pattern Recognition*, 120:108156, 2021. 1, 2

[18] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*, pages 1–21, 2022. 2

[19] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, pages 12993–13000, 2020. 3