This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Securing the Skies: A Comprehensive Survey on Anti-UAV Methods, Benchmarking, and Future Directions

Yifei Dong[1*†] Fengyi Wu[1*†] Sanjian Zhang[1*†] Guangyu Chen[1*†] Yuzhi Hu[1*†]
Masumi Yano[1†] Jingdong Sun[2] Siyu Huang[3] Feng Liu[4] Qi Dai[5] Zhi-Qi Cheng[1*‡]

[1]University of Washington [2]Carnegie Mellon University
[3]Clemson University [4]Drexel University [5]Microsoft Research

{sanjian, gychen, masumi76, zhiqics}@uw.edu, yd2616@columbia.edu, wufengyi98@gmail.com
hyz0929@bu.edu, jingdons@cs.cmu.edu, siyuh@clemson.edu, fl397@drexel.edu, qid@microsoft.com

Abstract

Unmanned Aerial Vehicles (UAVs) are indispensable for infrastructure inspection, surveillance, and related tasks, yet they also introduce critical security challenges. This survey provides a wide-ranging examination of the anti-UAV domain, centering on three core objectives—classification, detection, and tracking—while detailing emerging methodologies such as diffusion-based data synthesis, multi-modal fusion, vision-language modeling, self-supervised learning, and reinforcement learning. We systematically evaluate state-of-the-art solutions across both single-modality and multi-sensor pipelines (spanning RGB, infrared, audio, radar, and RF) and discuss large-scale as well as adversarially oriented benchmarks. Our analysis reveals persistent gaps in real-time performance, stealth detection, and swarm-based scenarios, underscoring pressing needs for robust, adaptive anti-UAV systems. By highlighting open research directions, we aim to foster innovation and guide the development of next-generation defense strategies in an era marked by the extensive use of UAVs.

1. Introduction

Unmanned Aerial Vehicles (UAVs) have expanding roles in civilian and military domains, offering benefits for infrastructure inspection, reconnaissance, and surveillance. Yet their rapid proliferation also introduces serious security threats, including unauthorized entry into restricted airspace and covert data gathering. As low-cost, agile drones become more prevalent, the need for robust counter-UAV (anti-UAV) technologies that can detect, track, and neutralize hostile UAVs grows increasingly urgent.

*Equal contribution. Student author order by lottery.
†Work done at UW Tacoma School of Engineering and Technology.
‡Corresponding author (zhiqics@uw.edu).

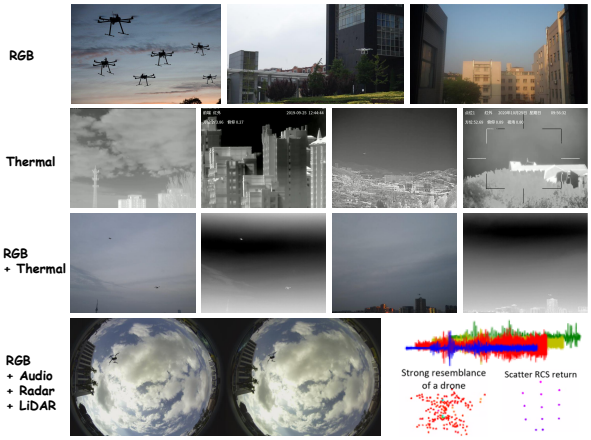


Figure 1. Datasets capturing UAV with diverse modalities: 1). RGB [163], 2). Thermal [51]; 3). RGB + Thermal [153]; 4). RGB + Audio + Radar + LiDAR [156].

Scope & Motivation. This survey provides a comprehensive overview of current anti-UAV research, focusing on three key objectives: *classification* (Section 3), *detection* (Section 4), and *tracking* (Section 5). We underscore the importance of leveraging multi-modal data—encompassing RGB, infrared (IR), radar, RF, and audio—to address small-object detection, erratic flight trajectories, and adversarial tactics. Figure 2 illustrates five primary research thrusts that drive modern anti-UAV innovations: (*i*) diffusion-based data synthesis and domain adaptation, (*ii*) vision-language modeling, (*iii*) self-supervised and unsupervised learning, (*iv*) multi-modal fusion, and (*v*) reinforcement learning.

Datasets & Benchmarking. The advancement of anti-UAV systems crucially depends on high-quality, large-scale datasets. Section 2 reviews pivotal benchmarks—including *DroneRF*, *Halmstad Drone*, and *Anti-UAV*—detailing their annotation fidelity, sampling rates, and environmental diversity. We also highlight the pressing need for more specialized datasets to cover scenarios such as night-time op-

Table 1. **Statistical Comparison of Public UAV Datasets.** "Seq." = number of sequences, "Frame" = number of frames, "T-Type" = target types, "S-Type" = scenario types. "Modality": RGB, IR, Audio (A), Radar (RA), LiDAR (L), RF. "Task": Classification (C), Detection (D), Tracking (T). "Annotation": class, mask, bbox, centroid, 3D coords (3D coordinates), or trajectory.

Dataset	Task	Modality	Seq.	Frame	Annotation	Resolution	Rate (Hz)	T-Type	S-Type	Venue (Year)
FL-Drones [103]	D, T	RGB	14	39K	Bbox	640×480 / 752×480	–	1	2	TPAMI (2016)
NPS-Drones [68]	D, T	RGB	50	70K	Bbox	1920×1280 / 1280×760	–	1	–	IROS (2016)
USC Drone [24]	D, T	RGB	60	20K	Bbox	1920×1080	15	1	–	APSIPA ASC (2017)
DroneRF [8]	C, D	RF	454	10M	Class	14-bit (RF)	Less than 1K (RF)	3	–	Data in Brief (2019)
DroneAudio [2]	C, D	A	–	1332 clips	Class	–	16K (A)	2	–	IWCMC (2019)
MAV-VID [100]	D, T	RGB	–	40K	Bbox	–	–	1	–	IEEE Access (2020)
Hui [55]	D, T	IR	22	16K	Centroid	256×256	100	1	22	CSD (2020)
Acoustic-UAV [21]	C, D	A	–	91K sec	Class	–	22.05K (A)	9	6	Drones (2021)
mDrone [165]	C, D	RA	100	60K	3D coords	–	10 (RA)	1	–	ICRA (2021)
Det-Fly [166]	D	RGB	–	13K	Bbox	3840×2160	5	1	4	RA-L (2021)
MOT-Fly [28]	C, D	RGB	16	11K	Bbox	1920×1080	–	3	–	ICUS (2021)
Halmstad Drone [122]	D, T	RGB, IR, A	650	203K	Bbox	640×512 (RGB), 320×256 (IR)	60 (IR), 30 (RGB)	4	3	ICPR (2021)
Drone-vs-Bird [29]	C, D	RGB	77	105K	Bbox	720×576 – 3840×2160	–	8	–	AVSS (2021)
Anti-UAV [58]	D, T	RGB, IR	318	297K	Bbox	1920×1080 (RGB), 640×512 (IR)	25	–	7	TMM (2021)
UAVSwarm [133]	D, T	RGB	72	12K	Bbox	446×276 – 1919×1079	–	19	13	Remote Sens. (2022)
Synthetic Drone [140]	C	RGB	–	2K	Class	256×256	–	4	10	CORD (2022)
DUT Anti-UAV [163]	D, T	RGB	20	35K	Bbox	160×240 – 3744×5616	–	1	7	T-ITS (2022)
AOT [10]	D, T	IR	4943	5.9M+	Bbox	2448×2048	10	–	–	AWS (2023)
Anti-UAV600 [171]	D, T	IR	600	723K	Bbox	640×512	25	1	–	Preprint (2023)
Anti-UAV410 [51]	D, T	IR	410	438K	Bbox	640×512	25	1	7	TPAMI (2024)
MMAUD [156]	C, D, T	RGB, A, RA, L	50	45K	Trajectory	2560×960 (RGB)	30 (RGB), 41.8 (A), 15 (RA), 10 (L)	6	–	ICRA (2024)
RGBT-Tiny [153]	D, T	RGB, IR	115	93K	Bbox	640×512	15	7	8	TPAMI (2025)

erations, swarm-based threats, and real-time interception. Table 1 contrasts these datasets, revealing persistent gaps in annotation standards, sensor synchronization, and adversarial realism, which limit standardized evaluation and hinder robust system development.

Classification, Detection, & Tracking. While detection locates UAVs in visual or sensor data, classification provides deeper insight into UAV types and flight profiles. In Section 3, we survey recent classification methods that integrate visual and non-visual features for improved accuracy and threat assessment. Section 4 examines state-of-the-art detection pipelines ranging from traditional feature-based to CNN- and transformer-based architectures, whereas Section 5 discusses leading tracking algorithms (e.g., YOLO-based, Siamese) designed to manage fast-moving UAVs in cluttered or adversarial settings.

Emerging Techniques & Future Directions. Section 6 spotlights key emerging paradigms—such as diffusion-driven augmentation, vision-language integration, self-supervised learning, advanced multi-modal fusion, and reinforcement learning for real-time control—and discusses their tremendous potential to reshape anti-UAV capabilities. Ultimately, these methods aim to enable adaptable, real-time systems equipped to handle evolving threats through domain adaptation, robust sensor integration, and rapid policy updates. Section 6 concludes with a forward-looking perspective on next-generation anti-UAV solutions, outlining how these emerging methodologies can drive future research and industrial innovation.

2. UAV Datasets & Benchmarking

Table 1 summarizes publicly available UAV datasets spanning various sensor modalities (e.g., RGB, IR, audio, radar, LiDAR, and RF), annotation schemes (e.g., bounding boxes, 3D coordinates), and sampling rates. This section highlights significant dataset initiatives, explores emerging trends, and underscores unresolved challenges.

Foundational Efforts & Modality Expansion. Initial benchmarks such as *FL-Drones* [103], *NPS-Drones* [68], and *USC Drone* [88, 138] provided core RGB datasets for aerial detection and tracking. Beyond the visual domain, *DroneRF* [8] captures electromagnetic signatures for RF-based UAV classification, while *DroneAudio* [2] and *Acoustic-UAV* [21] focus on acoustic features. More recent efforts extend coverage to multi-modal data: *Halmstad Drone* [120, 121] fuses RGB, IR, and audio, and *MMAUD* [156] integrates radar and LiDAR. Community-driven competitions (e.g., the *2nd* and *3rd Anti-UAV* Workshops [162, 164] and the *4th Anti-UAV Challenge* [89]) feature adversarial or cluttered scenarios that reflect real-world operational complexity, while *UAVSwarm* [133] and *mDrone* [165] emphasize large-scale swarm coordination or low-visibility tracking—both highlighting the need for robust multi-sensor fusion [102].

Complex Scenes & Modalities. Recent datasets target increasingly demanding environments and more granular annotations. For instance, *Anti-UAV600* [171] and *Anti-UAV410* [51] rely on infrared or thermal imaging to capture nighttime or harsh weather conditions, while *Drone-*

Table 2. **Representative UAV Classification Approaches by Modality.** *Disclaimer:* All entries summarize our own interpretation of each paper's methodology and reported outcomes. Since authors use distinct datasets, labels, and metrics, **no direct numerical comparison** is valid. *Abbrev.:* UAV = Unmanned Aerial Vehicle, CNN = Convolutional Neural Network, RNN = Recurrent Neural Network, SVM = Support Vector Machine, DNN = Deep Neural Network, GAN = Generative Adversarial Network, LSTM = Long Short-Term Memory, CLSTM = Contextual LSTM, DDI = Drone Detection and Identification, PSD = Power Spectral Density, MFCC = Mel-Frequency Cepstral Coefficients, LFCC = Linear Frequency Cepstral Coefficients, SCF = Spectral Correlation Functions, TWM = Time-frequency Waterfall Map, GTCC = GammaTone Cepstral Coefficients, SNR = Signal-Noise Ratio, APE = Average Pose Error, CV = Cross Validation, IR = Infrared, ID = Identification, Radar = RA, LiDAR = L.

Ref. (Yr.)	Mod.	Method	Dataset	Key Observations (Our Reading)
RF-Based				
Al-Sa'd et al. (2019) [3]	RF	3 DNNs (RF fingerprint)	DroneRF [8]	Introduced DroneRF (open-source, multiple flight modes). Developed three DNN classifiers for UAV presence, type, and flight mode. Achieved 99.7% (2-class), 84.5% (4-class), 46.8% (10-class). Forms an early baseline on DroneRF.
Al-Emadi et al. (2020) [1]	RF	CNN	DroneRF [8]	Achieved 99.8% (2-class), 85.8% (4-class), and 59.2% (10-class). Comparable to Al-Sa'd et al. [3] on 2-class tasks, showing CNN's viability for RF-based identification.
Medaiyese et al. (2021) [80]	RF	XGBoost pipeline	DroneRF [8]	Used 227 signal segments. Reached 99.96% (2-class), 90.73% (4-class), 70.09% (10-class). Overall stronger than Al-Sa'd et al. [3] for simpler tasks, but lags on higher-class classification vs. more recent solutions.
Li et al. (2021) [72]	RF	Bispectrum Siamese Network + Contrastive Learning	DroneRF [8]	Converts UAV signals to bispectrum for unsupervised representation. Achieved 100% (2-class), 98.57% (4-class), and 92.31% (10-class). Gains up to 10% over SimCLR, SimSiam, and GAN baselines, though still below XGBoost for large class sets.
Nemer et al. (2021) [86]	RF	Ensemble learning with hierarchical classification	DroneRF [8]	Multi-stage scheme yields 99% overall accuracy on ten-type classification. Outperforms earlier CNN approaches (e.g., [1]) for flight-mode tasks.
Inani et al. (2023) [56]	RF	XGBoost + PSD features + 1DCNN	DroneRF [8]	Achieved 100% (2-class), 99.82% (4-class), 99.51% (10-class)—current state-of-the-art on DroneRF. Improves upon [80] and [3] across all class splits, leveraging PSD feature fusion with deep learning.
Audio-Based				
Al-Emadi et al. (2019) [2]	Aud.	CNN, RNN, CRNN	DroneAudio [2]	Released DroneAudio dataset with $\sim$ 1300 drone clips. Demonstrated CNN-based acoustic ID, achieving 96.38% (2-class) and 92.94% (3-class) accuracy. Provided a baseline for audio-based UAV detection.
Katta et al. (2022) [61]	Aud.	CRNN, CNN, LSTM, CLSTM, Transformer	DroneAudio [2]	Benchmarked multiple neural architectures for UAV detection and ID on DroneAudio. LSTM outperformed CRNN/CNN with 98.93% accuracy (2-class) and 98.60% accuracy (3-class).
Xiao et al. (2025) [148]	Aud.	CNN+RNN (temporal)	MMAUD (audio-only) [156]	Achieved 98.0% classification accuracy and 0.55 APE for 3D trajectory estimation with a Mamba-based audio pipeline (4-channel short melspectrograms). Demonstrated robust performance across day/night conditions.
Radar-Based				
Mendis et al. (2016) [81]	Radar	CNN (micro-Dopp.)	None public	Utilized deep belief networks for micro-Doppler SCF signatures, identifying UAV presence within radar beam width. One of the earliest radar-based UAV classification approaches, but lacks open data for direct replication.
Vision-Based				
Wisniewski et al. (2022) [139]	RGB	DenseNet201 CNN	Synthetic Drone [140]	Achieved 92.4% classification accuracy (distinguishing DJI Phantom, Mavic, and Inspire) with 88.8% precision, 88.6% recall, 88.7% F1-score. Demonstrated synthetic data viability for UAV type classification.
Multi-Modal				
Deng et al. (2024) [32]	RGB, RA, L	Deep fusion	MMAUD [156]	Explored 4-class UAV classification in 3D pose estimation. Fused LiDAR, radar, and fisheye camera inputs via an EfficientNet-B7–based pipeline. Achieved 81.36% classification accuracy, improving keyframe-based multi-sensor approaches.
Xiao et al. (2024) [147]	RGB, Aud.	AV-DTEC (self-supervised audio-visual fusion)	MMAUD [156]	Achieved 99.3% classification accuracy and 0.67m APE. Parallel state-space fusion boosted performance under varied illumination. Offers an alternative to full multi-sensor pipelines by only focusing on audio-visual synergy.

vs-Bird [29] focuses on distinguishing UAVs from visually similar wildlife. Other repositories (e.g., *Midgard* [131] and *Rosner* [102]) explore indoor or GPS-denied settings, alongside multi-view or re-identification tasks [11, 59]. Simultaneously, annotations have expanded from bounding boxes to 3D coordinates and comprehensive trajectory information [156], aligning with the growing complexity of interception and multi-UAV maneuvers.

Trends & Open Challenges. Despite notable advances in dataset scope, several gaps persist. Collections like *MMAUD* [156] and *Halmstad Drone* [121] require precise sensor calibration and synchronization, yet standardized fusion protocols remain limited. While *UAVSwarm* [133] includes up to 23 drones, testing on larger fleets (50+) remains uncommon, constraining research on scaled cooperative tracking and interception. Moreover, stealth or jamming scenarios—emerging as key topics in anti-UAV forums [89, 162, 164]—are still sparsely represented in open-source datasets. High-resolution streams (e.g., *DetFly* [166] and *AOT* [10]) also challenge onboard systems in terms of storage and processing, prompting further research into compression and hardware acceleration [10, 11].

3. UAV Classification Approaches

Table 2 synthesizes representative UAV classification methods covering radio frequency (RF), audio, radar, vision, and LiDAR modalities. These methods range from simple drone-presence detection [87] to more nuanced classifica-

Table 3. **Representative UAV Detection Approaches by Modality.** *Disclaimer:* All entries summarize our own interpretation of each paper's methodology and reported outcomes. Since authors use distinct datasets, labels, and metrics, **no direct numerical comparison** is valid. *Abbrev.:* UAV = Unmanned Aerial Vehicle, CNN = Convolutional Neural Network, PSD = Power Spectral Density, LSTM = Long Short-Term Memory, DNN = Deep Neural Network, RNN = Recurrent Neural Network, YOLO = You Only Look Once, FMM = Feature Modeling Module, FCM = Feature Capture Module, LGFFM = Local-Global Feature Focusing Module, EMA = Efficient Multi-scale Attention, DCNv2 = Deformable Convolution v2, CSDE = Cross Stage Detail Enhance, CSP-PMSA = Cross Stage Partial, Partial Multi-scale, DERS = Detail Enhanced Rep Shared, MCJT = Multi-Consistency Joint Tracker, RMCM = Robust Motion Constraint Module, FSRM = Flexible Spatial Remapping Module, ATUS = Adaptive Template Update Strategy, MFCC = Mel Frequency Cepstrum Coefficient, RCNN = Region-based CNN, FPN = Feature Pyramid Network, SWIN = Shifted Window Transformer, DETR = Detection Transformer, OEN = Object Enhancement Net, IR = Infrared, LiDAR = L F1 = harmonic mean of precision & recall, mAP@0.5 = mean Average Precision at 0.5 IoU, mAP@0.5:0.95 = mean Average Precision at IoU in [0.5:0.95], AP@0.5 = Average Precision at 0.5 IoU, SOTA = state-of-the-art, RA = Radar.

Ref. (Yr.)	Mod.	Method	Dataset	Key Observations (Our Reading)
Radar-Based				
Zhao et al. (2022) [165]	RA	CNN + LSTM	mDrone [165]	Exploits Doppler signals from propellers for real-time detection and 3D tracking at 10 Hz. Achieves ≈9 cm localization error, outperforming five classic radar baselines (e.g., 3D MUSIC) by ≈66% in accuracy.
Vision-Based				
Seidaliyeva et al. (2020) [110]	RGB	Background subtraction + MobileNetV2	Drone-vs-Bird [29], BirdDetection [154]	Combines Drone-vs-Bird with a bird dataset to enhance UAV-vs-bird discrimination. Reached 0.56 F1 at IoU=0.5 (9 FPS). Performance is improved upon by works using more advanced YOLO variants (e.g., [31]).
Dadboud et al. (2021) [31]	RGB	YOLOv5x (PANet neck, mosaic augment)	Det-Fly [166], Drone-vs-Bird [29]	Combined Drone-vs-Bird + Det-Fly, achieving 0.98 mAP@0.5 with 0.96 recall. Improves upon [110] on Drone-vs-Bird, though different training splits limit exact comparison.
Sangam et al. (2023) [108]	RGB	CSPDarkNet-53 + VideoSwin	FL-Drones [103], NPS-Drones [68], AOT [10]	On FL-Drones: 0.89 precision, 0.85 recall, 0.87 F1, 0.84 AP@0.5. On NPS-Drones: 0.94 precision, 0.92 recall, 0.93 F1, 0.93 AP@0.5. On AOT: 0.85 precision, 0.76 recall, 0.80 F1, 0.82 AP@0.5. Comparable to Rebbapragada et al. [98] under similar setups.
Tu et al. (2024) [128]	RGB	YOLOv8 + FCM + FMM	DUT Anti-UAV[163]	AP@0.5=0.913, AP@0.5:0.95=0.612, below the best YOLO-based methods on this dataset (e.g., [53]) but highlights a lightweight, efficient design.
Cheng et al. (2024) [26]	RGB	YOLOv5s + LGFFM	DUT Anti-UAV [163]	mAP@0.5=0.943. Improves the recall rate of 2.96% and the precision of 1.41% compared with YOLOv5s. Slightly outperforms [128] on the same benchmark. Low-complexity approach with 4.4 ms/frame inference.
Bo et al. (2024) [16]	RGB	YOLOv7-tiny + Inception-NeXt + SPPFCSPC-SR + "Get-and-Send"	DUT Anti-UAV [163]	mAP@0.5=0.932 at 104 FPS. Balances fast inference and high accuracy, though slightly lower mAP@0.5 than [26] and [53].
Elsayed et al. (2024) [38]	RGB	YOLOv6 + LERFNet + large-kernel LAM	DUT Anti-UAV [163]	mAP@0.5=0.936, with 0.953 precision. Surpasses [16] in accuracy, albeit at presumably higher resource use than the YOLOv7-tiny approach.
Huang et al. (2024) [53]	RGB	YOLOv8 + ghost conv + EMA + DCNv2	Det-Fly [166], DUT Anti-UAV [163]	On DUT Anti-UAV, yields 0.971 mAP@0.5, highest among YOLO-based reports here ([16, 26, 38, 128]). On Det-Fly, obtains 0.934 mAP@0.5.
Sun et al. (2024) [118]	IR	YOLOv8 w/ triple-input, BiFormer attention, small-object layer	Hui [55], Anti-UAV [58], Anti-UAV410 [51]	On Hui dataset: 0.943 mAP@0.5 at 72 FPS. Generalizes well to Anti-UAV (0.948 mAP@0.5) and Anti-UAV410 (0.927 mAP@0.5). Demonstrates strong IR-based detection across multiple IR benchmarks.
Wisniewski et al. (2024) [141]	RGB	Faster R-CNN	MAV-VID [100], Anti-UAV [58], Drone-vs-Bird [29]	AP@0.5=0.970 on MAV-VID, AP@0.5=0.498 on Drone-vs-Bird, AP@0.5=0.678 on Anti-UAV.
Zhou et al. (2024) [169]	RGB	YOLOv8 + multi-scale spatial attention + HDC	Det-Fly [166]	Achieved 0.815 precision, 0.849 recall. Gains of +0.05 precision/+0.06 recall relative to YOLOv8 baseline. Lags behind [53] and [46] in absolute mAP on Det-Fly, but emphasizes robust small-object detection.
Rebbapragada et al. (2024) [98]	RGB	SWIN + DETR + OEN	FL-Drones [103], NPS-Drones [68], AOT [10]	Matches Sangam et al. [108] in precision/recall/AP@0.5 = 0.89/0.85/0.84 on FL-Drones and 0.94/0.92/0.93 on NPS-Drones with a transformer-based pipeline. On AOT dataset, obtains 0.85 precision, 0.76 recall, AP@0.5=0.82 comparable to [108].
Hao et al. (2025) [46]	RGB	YOLOv8 w/ CSDE, CSP-PMSA, OmniKernel, DERS head	Det-Fly [166], MOT-Fly [28]	On MOT-Fly, hits mAP@0.5=0.949 with only 1.13M parameters. On Det-Fly, achieves 0.964 mAP@0.5—slightly exceeding [53] at 0.934, though using different training splits.
Multi-Modal				
Svanström et al. (2021) [122]	RGB, IR, Aud.	YOLOv2 (visible) + MFCC (audio) + LSTM	Halmstad Drone [122]	Achieved F1≈0.61–0.88 across different distance with thermal IR sensor, F1≈0.69–0.86 with visible camera, F1≈0.93 with audio detector . Fusion cut false alarms vs. single-sensor methods, illustrating multi-modal benefits.
Wu et al. (2024) [143]	RGB, RA	Attention-based Spatiotemporal Fusion Net	None public	Achieved 93.0% precision, 83.7% recall, 88.7% mAP in occluded scenes at 20 FPS. Outperforms vision-only detectors by fusing radar signals, though dataset not publicly released.

tions of UAV types, flight modes, and orientations [109]. Although comparison is difficult due to inconsistent experimental protocols, modality-specific considerations—such as spectral noise or lighting variability—motivate distinct architectural and data-processing strategies. Below, we highlight key insights and emerging trends.

(1) RF-Based Classification. RF-focused approaches exploit the electromagnetic signatures emitted by UAVs to detect, identify, and even recognize specific flight modes. Al-Sa'd et al. [3] introduced the DroneRF dataset, demonstrating high accuracy (99.7%) on a 2-class task but lower performance (46.8%) on a more granular 10-class task. Subsequent work by Al-Emadi et al. [1] confirmed CNNs' viability for DroneRF, achieving 99.8% (2-class) and 59.2% (10-class). Recent studies have explored advanced feature engineering or ensemble learning: for instance, Nemer et al. [86] outperformed CNN-based baselines via hierarchical ensembles, while Li et al. [72] incorporated bispectrum representations with contrastive learning. Inani et al. [56] achieved near-perfect 10-class accuracy (99.51%) by fus-

ing PSD features, a 1D-CNN, and XGBoost. Despite these advancements, real-world deployments face challenges including spectral congestion, adversarial jamming, and domain shifts [9, 37, 132].

(2) Audio-Based Classification. Audio pipelines distinguish UAVs based on rotor noise and other acoustic signatures. Al-Emadi et al. [2] introduced the DroneAudio dataset, reporting 96.38% (2-class) and 92.94% (3-class) accuracy using CNNs. Katta et al. [61] subsequently demonstrated the benefits of LSTM, CRNN, and Transformers, with LSTM attaining 98.93% (2-class). Meanwhile, Xiao et al. [148] employed a temporal CNN+RNN to handle MMAUD's audio modality, reaching 98.0% accuracy alongside strong pose estimation results. Noise from wind, background chatter, or multiple UAVs remains an obstacle in practical scenarios [14, 57], motivating research into denoising filters and multi-microphone arrays [119].

(3) Radar-Based Classification. Radar offers micro-Doppler signatures that persist under reduced visibility or nighttime conditions. Early studies (e.g., Mendis et al. [81]) validated radar-based UAV classification using deep belief networks, though limited dataset availability inhibited broad adoption. CNN- and LSTM-based pipelines now dominate, focusing on micro-Doppler representations to improve robustness in cluttered environments [132, 166]. Nonetheless, radar hardware cost and calibration constraints remain practical challenges [132].

(4) Vision-Based Classification. Computer vision pipelines target UAV recognition via optical images, often focusing on manufacturer types (e.g., DJI Phantom, Mavic). Wisniewski et al. [139] achieved 92.4% accuracy on three drone types (Phantom, Mavic, Inspire) and illustrated synthetic imagery's viability [140]. Other studies address issues of small object size and motion blur, incorporating advanced detectors or domain adaptation [109, 130]. While performance in controlled scenarios is high, vision-based methods degrade under extreme illumination, adverse weather, or dense clutter [18, 33].

(5) Multi-Modal Recognition. Integrating multiple sensors (e.g., RF+Audio [43], Radar+LiDAR+Vision [156], or Audio+Vision [147]) can mitigate single-sensor blind spots. Deng et al. [32] fused fisheye RGB, radar, and LiDAR data from MMAUD, reaching 81.36% accuracy on a 4-class task, while Xiao et al. [147] combined audio and RGB streams to achieve 99.3% accuracy and sub-meter pose errors. Although multi-modal setups improve robustness, they also require careful calibration, synchronization, and greater computational resources [132, 156].

Key Insights & Challenges. Across these approaches [18, 33, 130, 132, 166], each modality faces inherent constraints: (1) *RF-based solutions* excel at flight-mode or type identification [3, 9], but radio congestion and potential adversarial attacks are critical hurdles [37, 132]. (2) *Audio-based methods* achieve high accuracy in controlled settings yet suffer from significant performance drops in high-noise or windy environments [2, 14, 57]. (3) *Radar-based pipelines* bypass visual occlusions but incur higher costs and calibration demands [81, 132]. (4) *Vision-based approaches* generally provide fine-grained recognition but degrade with lighting or weather changes [130, 139]. (5) *Multi-modal fusion* can mitigate single-sensor weaknesses and reduce false positives [32, 43] yet raises complexity in real-time deployment [156]. Additionally, adversarial robustness remains an open problem, particularly where sensors or neural networks may be fooled by spoofing or deceptive signals [61, 80]. A continuing research direction involves efficient, low-latency architectures [132, 148] that can handle on-board or edge-based processing for rapid UAV detection and classification.

4. UAV Detection Approaches

Table 3 summarizes UAV detection pipelines spanning radar, vision, and multi-modal solutions. Each modality addresses distinct obstacles—such as small object size, cluttered environments, and adverse weather—offering unique advantages and trade-offs. Although direct numerical comparisons are challenging due to varying datasets and evaluation protocols, these methods collectively illustrate emerging trends in detection algorithms and system design. We discuss salient insights from the literature below.

(1) Radar-Based Detection. Radar methods detect UAVs by exploiting micro-Doppler signatures, providing resilience under low-visibility or harsh weather conditions [132, 165]. For instance, Zhao et al. [165] achieved decimeter-level localization at 10 Hz by combining CNNs with LSTMs, outperforming conventional radar baselines. However, radar sensors can suffer from interference and clutter, particularly in urban areas or heavily congested electromagnetic environments [143].

(2) Vision-Based Detection. Vision-based techniques remain widely adopted, benefiting from readily available cameras and sophisticated CNN-based detectors [124, 170]. Early work centered on region-based methods [99] and YOLO architectures [17, 135], gradually evolving into more lightweight or attention-oriented variants [26, 46, 53]. While mosaic augmentation and specialized modules can improve small-object detection [26, 31], vision-based approaches remain vulnerable to occlusions, clutter, and extreme distances. Researchers have thus explored thermal and infrared (IR) imaging [85, 118] or speed-accuracy trade-offs for real-time embedded scenarios [16, 38].

(3) Multi-Modal Fusion. Single-modality approaches often fail under suboptimal conditions—low-visibility (vision), interference (radar), or noise (audio) [2, 129]. Consequently, multi-modal fusion has gained traction as a means of leveraging complementary sensor data [12, 120]. Svanström et al. [120] showed that augmenting YOLO-

Table 4. **Representative UAV Tracking Approaches.** *Disclaimer:* All entries summarize our own interpretation of each paper's methodology and reported outcomes. Since authors use distinct datasets, labels, and metrics, **no direct numerical comparison** is valid. *Abbrev.:* UAV = Unmanned Aerial Vehicle, CNN = Convolutional Neural Network, RNN = Recurrent Neural Network, SVM = Support Vector Machine, DNN = Deep Neural Network, LSTM = Long Short-Term Memory, GMM = Gaussian Mixture Model, IR = Infrared, TIR = Thermal Infrared, ID = Identification, Radar = RA, LiDAR = L, SOTA = State of the Art, YOLO = You Only Look Once, YOLOX = You Only Look Once Extended, DFSC = Dual-Flow Semantic Consistency, LoFTR = Detector-Free Local Feature Matching with Transformers, SiamRPN++ = Siamese Region Proposal Network++, SiamSTA = Siamese Spatio-Temporal Attention tracker, SiamSRT = Siamese Search Region-free Tracker, SiamDT = Siamese Dual-semantic Tracker, SiamFusion = Siamese Modality-Fusion Tracker, DCF = Discriminative Correlation Filter, CDCF = Change Detection-based Correlation Filter, CFRM = Channel Feature Refinement Module, BYTE = Bounding-box Embedding Tracker with Association, STG-RPN = Spatio-Temporal Guided Region Proposal Network, ID-RCNN = Instance Discrimination Region-based CNN, FPS = Frames Per Second, Acc = Accuracy, F1 = F1 Score, TLD = Tracking-Learning-Detection, DFSC = Dual-Flow Semantic Consistency, ATUS = Adaptive Template Updating Strategy, EECF = Eagle-Eye Correlation Filter, ECA = Efficient Channel Attention, STFTrack = Spatio-Temporal Focused Tracker, CAM = Channel Attention Module, MCJT = Multi-Consistency Joint Tracker, α-IoU = Alpha Intersection over Union, Swin = Shifted Window Transformer, MSE = Mean Squared Error, SPLT = Self-paced Learning Tracker.

Ref. (Yr.)	Mod.	Method	Dataset	Key Observations (Our Reading)
Detection-Based Trackers				
Jiang et al. (2023) [58]	RGB, IR	GlobalTrack + DFSC	Anti-UAV [58]	Released 318-sequence RGB-T UAV tracking benchmark. Achieved the best state accuracy score of 74.33% on the validation set and 65.41% on the test set
Wu et al. (2024) [144]	RGB	DCF + dual-fovea + contrast + opponent-color (EECF)	Anti-UAV [58], USC-Drone [24]	Real-time EECF tracker (86 FPS) via dual-fovea design. Achieved the state accuracy score of 66.42% on Anti-UAV, 70.45% on USC-Drone.
Wang et al. (2024) [134]	RGB	YOLOX + Swin-T + α-IoU + BYTE	UAVSwarm [133]	Detection-driven approach with α-IoU + ByteTrack. Achieved 78.2% tracking accuracy on UAVSwarm test split.
Deng et al.(2024) [32]	RGB, RA, L	YOLOv9 + EfficientNet-B7 + LSTM	MMAUD [156]	Intergrated YOLOv9 + EfficientNet-B7 + LSTM/Kalman for classification and tracking. Achieved 2.21 Pose MSE Loss on 3D tracking of MMAUD and ranked 1st in UG2+ Challenge.
Siamese (Contrastive) Trackers				
Huang et al. (2021) [49]	IR	SiamSTA	1st / 2nd Anti-UAV Challenges [58] [162]	Achieved 74.46 and 67.30 scores on 1st / 2nd Anti-UAV test-dev sets, ranking 1st. Extended SiamSTA with spatio-temporal attention + motion estimation.
Li et al. (2022) [71]	IR	SuperDiMP (ResNet50)	2nd Anti-UAV Challenge [162]	Used SuperDiMP with dual-branch IR fusion. Achieved +5.84% SR and +4.52% PR over baseline. Focused on contrast inversion robustness.
Chen et al. (2022) [22]	IR	SiamSTA + CDCF	1st Anti-UAV Challenge [58]	SiamSTA with CDCF and 3-stage redetection. Achieved top precision in 2nd Anti-UAV Challenge. Achieved 68.26 and 76.11 scores on Anti-UAV test and validation sets. Robust on fast/small UAVs.
Zhang et al. (2023) [161]	RGB, IR	SiamFusion	Anti-UAV [58]	Achieved overall 63.37 score on Anti-UAV with 14 FPS. Developed a dual-fusion tracker with modality + decision-level fusion. Used Swin backbone + Local-Global Converter.
Xie et al. (2023) [149]	IR	FPN-based Siamese + STG-RPN + ID-RCNN (STF-Track)	Anti-UAV [58]	STFTrack using STG-RPN and ID-RCNN. Achieved 91.2% precision, 66.6% success @ 12.4 FPS. Addressed distractors and thermal crossover.
Zhang et al. (2024) [158]	IR	OSTrack + MCJT (RMCM + FSRM + ATUS)	1st / 2nd / 3rd Anti-UAV Challenges [58] [162] [164]	Achieved 69.65% Acc, 80.45% Acc, 65.88% Acc and 50.28% Acc on the 1st Anti-UAV test set, 1st Anti-UAV test-dev, 2nd Anti-UAV test-dev, and 3rd Anti-UAV validation set. Proposed MCJT (RMCM + FSRM + ATUS) with adaptive template and multi-consistency. Ran at 44 FPS.
Huang et al. (2024) [50]	IR	SiamSRT + C-C RPN + S-L RCNN + TMB + SCFD	Anti-UAV [58]	Achieved 71.64 and 80.04 scores on Anti-UAV testing and validation sets. Global Siamese tracker using Swin-T + consistency modules (TMB + SCFD). Avoided template degradation; outperformed multiple SOTA trackers.
Huang et al. (2024) [51]	IR	SiamDT (Swin Transformer + Dual-Semantic RPN + Versatile R-CNN)	Anti-UAV410 [51]	Introduced Anti-UAV410 dataset. Proposed SiamDT with dual semantic reasoning + background suppression. Achieved 68.19 and 71.65 scores on its testing and validation sets.
Hybrid (Detection + Siamese) Trackers				
Cheng et al. (2022) [25]	IR	Multi-Hybrid Attention + SiamRPN++ + YOLOv5 (SiamAD)	Anti-UAV [58]	SiamAD with CSAM/CAM + hierarchical discriminator + YOLOv5. 88.4% DP @ 37.1 FPS. Improved long-term IR tracking.
Yu et al. (2023) [155]	IR	Transformer + YOLOv5 + LoFTR + GMM + OSTrack	1st / 2nd Anti-UAV Challenges [58] [162]	Achieved 2nd place in Anti-UAV Challenge (3rd) with 68.8 score. And 75.13% and 80.38% accuracy on the 1st and 2nd test dev, individually. Integrated OS-Track, YOLOv5, LoFTR, GMM for multi-stage IR tracking.

based RGB or thermal imaging with audio features increased F1 scores and reduced false alarms. Similarly, Wu et al. [143] fused radar and RGB streams via attention-based spatiotemporal networks to improve occlusion handling. Practical adoption, however, demands meticulous sensor synchronization, balanced hardware costs, robust computational efficiency, and real-time data association [98].

Key Insights & Challenges. From prior work [124, 132, 170], several recurring themes emerge. *1) Small UAV Detection:* Limited pixel coverage and subtle micro-Doppler signatures complicate detection [20, 47, 77], prompting multi-scale feature extraction, deformable convolutions, and transformer backbones [26, 53, 124]. Ensuring real-time operation at high accuracy remains challenging [132, 135]. *2) Environmental Factors:* Ambient noise (audio), lighting (vision), and radio interference (radar) degrade performance [2, 113, 129, 165], highlighting the need for advanced clutter suppression and robust sensor hardware. *3) Generalization vs. Specialization:* YOLO-based models offer generalizable frameworks [26, 31, 46], but specialized domains (e.g., stealth UAVs) may require niche architectures or domain-adaptive strategies [16, 143]. *4) Multi-Modal Fusion:* Complementary sensors effectively reduce false positives and enhance accuracy [120, 143], although issues like sensor alignment, real-time latency, and hardware cost must be resolved [12, 98]. *5) Benchmarking & Deployment:* Heterogeneous data sources, labeling formats, and test conditions [35, 93, 170] complicate standard-

ized evaluations. Though laboratory demonstrations show promise, real-time performance in crowded outdoor settings remains largely underexplored [39, 42]. Overcoming these challenges will be crucial for field-ready anti-UAV systems.

5. Anti-UAV Tracking Approaches

Table 4 reviews prominent UAV tracking methods spanning detection-based, Siamese-based, and hybrid trackers, employing RGB, IR, and multi-modal inputs. Although direct numerical comparisons are hampered by dataset and metric variations, these approaches collectively represent the state of the art in UAV tracking.

(1) Detection-Based Trackers. Detection-based pipelines usually prioritize real-time performance and utilize efficient object detectors. For example, Wu et al. [144] proposed EECF, a dual-fovea correlation filter tracker operating at 86 FPS with a state accuracy score of 66.42% on Anti-UAV [58] and 70.45% on USC-Drone [24]. Wang et al. [134] integrated YOLOX [44] with Swin Transformer [77] and ByteTrack, achieving 78.2% accuracy on UAVSwarm [133]. Deng et al. [32] combined YOLOv9 [136] and EfficientNet-B7 [123] with LSTM/Kalman filtering for 3D tracking, earning top-ranked pose MSE on MMAUD [156] and winning the UG2+ Challenge [90].

(2) Siamese (Contrastive) Trackers. Siamese trackers excel in learning robust feature similarities and handling fast or evasive UAV motion. Huang et al. [49] claimed first-place in the 1st and 2nd Anti-UAV Challenges [58, 162] with SiamSTA, which incorporates sophisticated spatio-temporal attention and motion estimation. Li et al. [71] augmented SuperDiMP with a dual-branch IR fusion strategy, boosting precision by 4.52% in the 2nd Anti-UAV Challenge. Chen et al. [22] further enhanced SiamSTA with CDCF and a multi-stage re-detection mechanism, excelling against small, fast UAVs. Beyond single-modality, Zhang et al. [161]'s SiamFusion seamlessly integrated RGB and IR streams for a 63.37 score on Anti-UAV, illustrating the tangible value of complementary sensor modalities.

(3) Hybrid (Detection + Siamese) Trackers. Hybrid methods combine detection pipelines with Siamese matching, capitalizing on each paradigm's respective strengths. Cheng et al. [25] introduced SiamAD, a synthesis of YOLOv5, multi-hybrid attention modules, and SiamRPN++, achieving 88.4% detection precision at 37.1 FPS on the Anti-UAV dataset [58]. Yu et al. [155] integrated Transformers, YOLOv5, LoFTR, GMM, and OSTrack for second place in the 3rd Anti-UAV Challenge [164], affirming the practical effectiveness of multi-stage IR-based tracking.

Key Insights & Challenges. Although real-time Siamese trackers [111] and detection-based pipelines [40] can sustain high frame rates, they often struggle under clutter, occlusion, or swarm conditions, thereby motivating more adaptive architectures [23, 150]. Stealth maneuvers and jamming further complicate UAV tracking, emphasizing the necessity for robust domain adaptation and dynamic cue prioritization [34]. While RGBT sensor fusion [67, 125] mitigates individual sensor shortcomings, incorporating LiDAR or radar [32] inevitably increases calibration overhead and computational load. Existing benchmarks rarely capture large-scale swarms, diverse weather conditions, or severe occlusions, consequently restricting methodical progress [145]. Transformer-based solutions [23, 30, 146] address complex motion patterns and partial occlusion but can lag in real-time performance, necessitating specialized hardware optimizations for on-board or edge devices. Extending multi-object trackers (e.g., ByteTrack [159]) to larger UAV swarms remains challenging due to association uncertainty and domain-specific constraints. Ultimately, next-generation trackers will demand faster inference, improved sensor fusion, and broader datasets to achieve reliable anti-UAV performance in real-world conditions.

6. Emerging Techniques & Future Directions

Recent UAV research has embraced advanced machine learning paradigms to boost performance, scalability, and autonomy. Figure 2 categorizes these innovations into five key themes: (*i*) diffusion models for data synthesis and domain adaptation, (*ii*) vision-language models for semantic understanding, (*iii*) self-/unsupervised learning, (*iv*) multi-modal fusion and sensor integration, and (*v*) reinforcement learning for autonomous control. We outline critical developments and future opportunities below.

Diffusion Models for Data Synthesis & Domain Adaptation. Diffusion models [48, 116] generate high-fidelity UAV imagery to address data scarcity and enhance domain diversity, including text-conditioned approaches [97, 105]. Recent advances in faster sampling [95, 106] support near real-time augmentation, while diffusion-based domain adaptation [13, 94, 160] narrows the gap between synthetic and real-world data. Future research may focus on on-the-fly adaptation from live sensor streams, enabling continuous refinement under dynamic UAV operating conditions.

Vision-Language Models & Semantic Context. Vision-language (VL) methods integrate textual cues with visual features to facilitate open-vocabulary UAV detection and classification [45, 96, 157, 167]. UAV-specific adaptations [52, 73, 76] incorporate textual descriptions on flight profiles, mission goals, or payload attributes, thereby *significantly* enriching threat assessment. By leveraging semantic narratives (e.g., infiltration routes, suspect cargo), future VL-based systems can *swiftly* adapt to novel UAV categories and enhance situational awareness.

Self-Supervised & Unsupervised Learning. Self- and unsupervised approaches alleviate dependence on labeled aerial data across varying altitudes, sensors, and weather [64, 65, 152]. Recent work addresses swarm-wide gener-

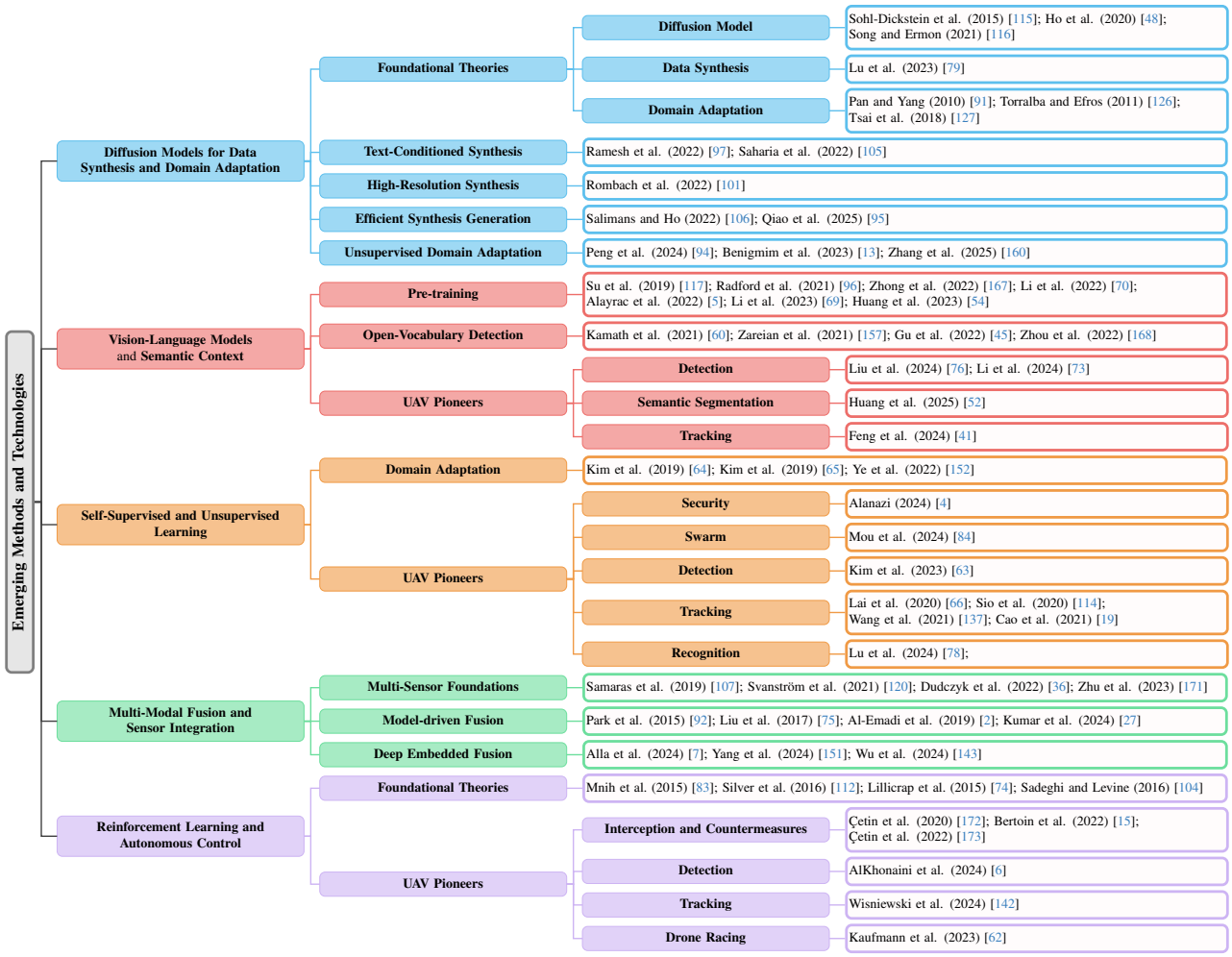


Figure 2. Taxonomy for emerging methods and technologies in UAV research.

alization [84], *further* strengthening multi-UAV detection [63], tracking [66, 137], and recognition [114], as well as security [4]. Moving forward, large-scale self-supervised pre-training that *rapidly* adapts to domain shifts in real time may enable *ongoing* responses to evolving UAV behaviors and stealth tactics.

Multi-Modal Fusion & Sensor Integration. Combining sensors (e.g., RGB, IR, radar, and audio) *significantly* boosts UAV detection and tracking robustness [82, 92, 107, 171]. Classical fusion [75, 92] is now augmented by deep embedding methods [7, 151], allowing adaptive sensor weighting under real-time constraints. However, *maintaining* precise synchronization [143] and carefully balancing power or bandwidth remain essential considerations for *robust*, large-scale deployments.

Reinforcement Learning & Autonomous Control. Reinforcement Learning (RL) has proven effective for adaptive UAV decision-making in dynamic environments, building on foundational breakthroughs [83, 112]. Modern UAV-centric RL research tackles interception [15, 173], detection [6], tracking [142], and high-speed racing [62], frequently integrating generative or multi-modal inputs for rapid policy updates. Future directions may emphasize hierarchical or multi-agent RL for large-scale swarm control, balancing cooperative behavior with adversarial countermeasures and mission-specific constraints.

7. Conclusion

This survey examined the evolving anti-UAV landscape, focusing on detection, tracking, and recognition. We highlighted recent progress in sensor fusion, deep learning, and multi-modal systems, emphasizing the urgency of robust, real-time solutions under adverse conditions. Analysis of representative datasets revealed gaps in adversarial robustness and scalable swarm tracking. Key insights underscore cross-modal integration, resilience to adversarial tactics, and the promise of self-supervised and reinforcement learning. Moving forward, generative modeling, semantic understanding, and adaptive multi-modal frameworks are poised to drive the next wave of anti-UAV innovations.

References

[1] Sara Al-Emadi and Felwa Al-Senaid. Drone detection approach based on radio-frequency using convolutional neural network. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 29–34, 2020. 3, 4

[2] Sara Al-Emadi, Abdulla Al-Ali, Amr Mohammad, and Abdulaziz Al-Ali. Audio based drone detection and identification using deep learning. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 459–464, 2019. 2, 3, 5, 6, 8

[3] Mohammad F Al-Sa'd, Abdulla Al-Ali, Amr Mohamed, Tamer Khattab, and Aiman Erbad. Rf-based drone detection and identification using deep learning approaches: An initiative towards a large open source drone database. *Future Generation Computer Systems*, 100:86–97, 2019. 3, 4, 5

[4] Abed Alanazi. Ssrl-uavs: A self-supervised deep representation learning approach for gps spoofing attack detection in small unmanned aerial vehicles. *Drones*, 8(9), 2024. 8

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022. 8

[6] Arwa AlKhonaini, Tarek Sheltami, Ashraf Mahmoud, and Muhammad Imam. Uav detection using reinforcement learning. *Sensors*, 24(6), 2024. 8

[7] Ildi Alla, Hervé B. Olou, Valeria Loscri, and Marco Levorato. From sound to sight: Audio-visual fusion and deep learning for drone detection. In *Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, page 123–133, New York, NY, USA, 2024. Association for Computing Machinery. 8

[8] MHD Saria Allahham, Mohammad F Al-Sa'd, Abdulla Al-Ali, Amr Mohamed, Tamer Khattab, and Aiman Erbad. Dronerf dataset: A dataset of drones for rf-based detection, classification and identification. *Data in brief*, 26:104313, 2019. 2, 3

[9] Mhd Saria Allahham, Tamer Khattab, and Amr Mohamed. Deep learning for rf-based drone detection and identification: A multi-channel 1-d convolutional neural networks approach. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 112–117. IEEE, 2020. 5

[10] Amazon. Airborne object tracking dataset. Registry of Open Data on AWS, 2023. Available at: https://registry.opendata.aws/airborne-object-tracking. 2, 3, 4

[11] Josh Andle, Nicholas Soucy, Simon Socolow, and Salimeh Yasaei Sekeh. The stanford drone dataset is more complex than we think: An analysis of key characteristics. *IEEE Transactions on Intelligent Vehicles*, 8(2):1863–1873, 2022. 3

[12] Anastasios Arsenos, Evangelos Petrongonas, Orfeas Filippopoulos, Christos Skliros, Dimitrios Kollias, and Stefanos Kollias. Nefeli: A deep-learning detection and tracking pipeline for enhancing autonomy in advanced air mobility. *Aerospace Science and Technology*, 155:109613, 2024. 5, 6

[13] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. One-shot unsupervised domain adaptation with personalized diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 698–708, 2023. 7, 8

[14] Andrea Bernardini, Federica Mangiatordi, Emiliano Pallotti, and Licia Capodiferro. Drone detection by acoustic signature identification. *electronic imaging*, 29:60–64, 2017. 5

[15] David Bertoin, Adrien Gauffriau, Damien Grasset, and Jayant Sen Gupta. Autonomous drone interception with deep reinforcement learning. In *12th International Workshop on Agents in Traffic and Transportation (ATT 2022) in conjunction with IJCAI-ECAI 2022*, number 3173, 2022. 8

[16] Chunjuan Bo, Yuntao Wei, Xiujia Wang, Zhan Shi, and Ying Xiao. Vision-based anti-uav detection based on yolov7-gs in complex backgrounds. *Drones*, 8(7), 2024. 4, 5, 6

[17] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[18] Aldrich A Cabrera-Ponce, J Martinez-Carranza, and Caleb Rascon. Detection of nearby uavs using cnn and spectrograms. In *Proceedings of the International Micro Air Vehicle Conference and Competition (IMAV)(Madrid), Madrid, Spain*, 2019. 5

[19] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15457–15466, 2021. 8

[20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 6

[21] Pietro Casabianca and Yu Zhang. Acoustic-based uav detection using late fusion of deep neural networks. *Drones*, 5(3), 2021. 2

[22] Junjie Chen, Bo Huang, Jianan Li, Ying Wang, Moxuan Ren, and Tingfa Xu. Learning spatio-temporal attention based siamese network for tracking uavs in the wild. *Remote Sensing*, 14(8):1797, 2022. 6, 7

[23] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 7

[24] Yueru Chen, Pranav Aggarwal, Jongmoo Choi, and C-C Jay Kuo. A deep learning approach to drone monitoring. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 686–691. IEEE, 2017. 2, 6, 7

[25] Feng Cheng, Zhiqiang Liang, Guoping Peng, Shun Liu, Shijie Li, and Ming Ji. An anti-uav long-term tracking

method with hybrid attention mechanism and hierarchical discriminator. *Sensors*, 22(10):3701, 2022. 6, 7

[26] Qian Cheng, Jia Li, Juan Du, and Shaojuan Li. Anti-uav detection method based on local-global feature focusing module. In *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1413–1418, 2024. 4, 5, 6

[27] Amit Kumar Chiranjeevi and Ozkan Giridhar. UAV detection multi-sensor data fusion. *Journal of Research in Science and Engineering*, 6(7):6–12, 2024. 8

[28] Zhaochen Chu, Tao Song, Ren Jin, and Tao Jiang. An experimental evaluation based on new air-to-air multi-uav tracking dataset. In *2023 IEEE International Conference on Unmanned Systems (ICUS)*, pages 671–676. IEEE, 2023. 2, 4

[29] Angelo Coluccia, Andrea Fascista, Arne Schumann, Lukas Sommer, Anastasios Dimou, Dimitrios Zarpalas, and et al. Drone-vs-bird detection challenge at IEEE avss 2021. In *17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021. 2, 3, 4

[30] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022. 7

[31] Fardad Dadboud, Vaibhav Patel, Varun Mehta, Miodrag Bolic, and Iraj Mantegh. Single-stage uav detection and classification with yolov5: Mosaic data augmentation and panet. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021. 4, 5, 6

[32] Tianchen Deng, Yi Zhou, Wenhua Wu, Mingrui Li, Jingwei Huang, Shuhong Liu, Yanzeng Song, Hao Zuo, Yanbo Wang, Yutao Yue, Hesheng Wang, and Weidong Chen. Multi-modal uav detection, classification and tracking algorithm – technical report for cvpr 2024 ug2 challenge, 2024. 3, 5, 6, 7

[33] Vedanshu Dewangan, Aditya Saxena, Rahul Thakur, and Shrivishal Tripathi. Application of image processing techniques for uav detection using deep learning and distance-wise analysis. *Drones*, 7(3):174, 2023. 5

[34] Nhat-Tan Do, Nhi Ngoc-Yen Nguyen, Dieu-Phuong Nguyen, and Trong-Hop Do. Ramots: A real-time system for aerial multi-object tracking based on deep learning and big data technology. *arXiv preprint arXiv:2502.03760*, 2025. 7

[35] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 6

[36] Janusz Dudczyk, Roman Czyba, and Krzysztof Skrzypczyk. Multi-sensory data fusion in terms of UAV detection in 3d space. *Sensors*, 22(12):4323, 2022. 8

[37] Ibrahim Elleuch, Ali Pourranjbar, and Georges Kaddoum. Leveraging transformer models for anti-jamming in heavily attacked uav environments. *IEEE Open Journal of the Communications Society*, 5:5337–5347, 2024. 5

[38] Mohamed Elsayed, Mohamed Reda, Ahmed S. Mashaly, and Ahmed S. Amein. Lerfnet: an enlarged effective receptive field backbone network for enhancing visual drone detection. *Vis. Comput.*, 41:2219–2232, 2024. 4, 5

[39] Jiwei Fan, Xiaogang Yang, Ruitao Lu, Xueli Xie, and Weipeng Li. Design and implementation of intelligent inspection and alarm flight system for epidemic prevention. *Drones*, 2021. 7

[40] Houzhang Fang, Xiaolin Wang, Zikai Liao, Yi Chang, and Luxin Yan. A real-time anti-distractor infrared uav tracker with channel feature refinement module. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1240–1248, 2021. 7

[41] Xiaokun Feng, Xuchen Li, Shiyu Hu, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Memvlt: Vision-language tracking with adaptive memory-based prompts. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. 8

[42] Timofey Filkin, Natalia Sliusar, Marco Ritzkowski, and Marion Huber-Humer. Unmanned aerial vehicles for operational monitoring of landfills. *Drones*, 2021. 7

[43] Alan Frid, Yehuda Ben-Shimol, Erez Manor, and Shlomo Greenberg. Drones detection using a fusion of rf and acoustic features and deep neural networks. *Sensors*, 24(8), 2024. 5

[44] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021, 2021. 7

[45] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022. 7, 8

[46] Hexiang Hao, Yueping Peng, Zecong Ye, Baixuan Han, Xuekai Zhang, Wei Tang, Wenchao Kang, and Qilong Li. A high performance air-to-air unmanned aerial vehicle target detection model. *Drones*, 9(2), 2025. 4, 5, 6

[47] Xin He, Kuangang Fan, and Zhitao Xu. Uav identification based on improved yolov7 under foggy condition. *Signal, Image and Video Processing*, 18(8):6173–6183, 2024. 6

[48] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7, 8

[49] Bo Huang, Junjie Chen, Tingfa Xu, Ying Wang, Shenwang Jiang, Yuncheng Wang, Lei Wang, and Jianan Li. Siamsta: Spatio-temporal attention based siamese tracker for tracking uavs. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1204–1212, 2021. 6, 7

[50] Bo Huang, Zeyang Dou, Junjie Chen, Jianan Li, Ning Shen, Ying Wang, and Tingfa Xu. Searching region-free and template-free siamese network for tracking drones in tir videos. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2023. 6

[51] Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Zhao, and Tingfa Xu. Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the

wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):2852–2865, 2024. 1, 2, 4, 6

[52] Bangju Huang, Junhui Li, Wuyang Luan, Jintao Tan, Chenglong Li, and Longyang Huang. Expanding open-vocabulary understanding for uav aerial imagery: A vision–language framework to semantic segmentation. *Drones*, 9 (2):155, 2025. 7, 8

[53] Min Huang, Wenkai Mi, and Yuming Wang. Edgs-yolov8: An improved yolov8 lightweight uav detection model. *Drones*, 8(7), 2024. 4, 5, 6

[54] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Bjorck, Vishrav Chaudhary, Subhojit Som, XIA SONG, and Furu Wei. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*, pages 72096–72109. Curran Associates, Inc., 2023. 8

[55] Bingwei Hui, Zhiyong Song, Hongqi Fan, Ping Zhong, Weidong Hu, Xiaofeng Zhang, Jianguo Lin, Hongyan Su, Wei Jin, Yongjie Zhang, and Yaxi Bai. A dataset for infrared image dim-small aircraft target detection and tracking under ground / air background, 2019. 2, 4

[56] Kalit Naresh Inani, K.S. Sangwan, and Dhiraj. Machine learning based framework for drone detection and identification using rf signals. In *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–8, 2023. 3, 4

[57] Sungho Jeon, Jong-Woo Shin, Young-Jun Lee, Woong-Hee Kim, YoungHyoun Kwon, and Hae-Yong Yang. Empirical study of drone sound detection in real-life environment with deep neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1858–1862, 2017. 5

[58] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Guodong Guo, Qixiang Ye, Jianbin Jiao, Jian Zhao, and Zhenjun Han. Anti-uav: A large-scale benchmark for vision-based uav tracking. *IEEE Transactions on Multimedia*, 25:486–500, 2023. 2, 4, 6, 7

[59] Isha Kalra, Maneet Singh, Shruti Nagpal, Richa Singh, Mayank Vatsa, and PB Sujit. Dronesurf: Benchmark dataset for drone-based face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 3

[60] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021. 8

[61] Sai Srinadhu Katta, Sivaprasad Nandyala, Eduardo Kugler Viegas, and Abdelrahman AlMahmoud. Benchmarking audio-based deep learning models for detection and identification of unmanned aerial vehicles. In *2022 Workshop on Benchmarking Cyber-Physical Systems and Internet of Things (CPS-IoTBench)*, pages 7–11, 2022. 3, 5

[62] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023. 8

[63] Juann Kim, Mia Y. Wang, and Eric T. Matson. Self-supervised drone detection using acoustic data. In *2023 Seventh IEEE International Conference on Robotic Computing (IRC)*, pages 67–70, 2023. 8

[64] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7, 8

[65] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12456–12465, 2019. 7, 8

[66] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6478–6487, 2020. 8

[67] Jing Li, Dong Hye Ye, Timothy Chung, Mathias Kolsch, Juan Wachs, and Charles Bouman. Multi-target detection and tracking from a single camera in unmanned aerial vehicles (uavs). In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4992–4997, 2016. 7

[68] Jing Li, Dong Hye Ye, Timothy Chung, Mathias Kolsch, Juan Wachs, and Charles Bouman. Multi-target detection and tracking from a single camera in unmanned aerial vehicles (uavs). In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4992–4997, 2016. 2, 4

[69] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 8

[70] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022. 8

[71] Shaogang Li, Jin Gao, Liang Li, Gang Wang, Yizheng Wang, and Xin Yang. Dual-branch approach for tracking uavs with the infrared and inverted infrared image. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1803–1806, 2022. 6, 7

[72] Taotao Li, Zhen Hong, Qianming Cai, Li Yu, Zhenyu Wen, and Renyu Yang. Bissiam: Bispectrum siamese network based contrastive learning for uav anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12109–12124, 2023. 3, 4

[73] Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. Toward open vocabulary aerial

object detection with clip-activated student-teacher learning. In *Computer Vision – ECCV 2024*, pages 431–448, Cham, 2025. Springer Nature Switzerland. 7, 8

[74] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019. 8

[75] Hao Liu, Zhiqiang Wei, Yitong Chen, Jie Pan, Le Lin, and Yunfang Ren. Drone detection based on an audio-assisted camera array. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 402–406, 2017. 8

[76] Jie Liu, Jinzong Cui, Mao Ye, Xiatian Zhu, and Song Tang. Shooting condition insensitive unmanned aerial vehicle object detection. *Expert Systems with Applications*, 246:123221, 2024. 7, 8

[77] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 6, 7

[78] Gejiacheng Lu, Xue Fu, Juzhen Wang, Hao Huang, Yu Wang, Yun Lin, and Guan Gui. A novel semi-supervised learning method using self-adaptive threshold for uav recognition. In *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, pages 1–5, 2024. 8

[79] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023. 8

[80] Olusiji O Medaiyese, Abbas Syed, and Adrian P Lauf. Machine learning framework for rf-based drone detection and identification system. In *2021 2nd International Conference On Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS)*, pages 58–64. IEEE, 2021. 3, 5

[81] Gihan J. Mendis, Tharindu Randeny, Jin Wei, and Arjuna Madanayake. Deep learning based doppler radar for micro uas detection and classification. In *MILCOM 2016 - 2016 IEEE Military Communications Conference*, pages 924–929, 2016. 3, 5

[82] Marco Messina and Gianpaolo Pinelli. Classification of drones with a surveillance radar signal. In *Computer Vision Systems – 12th International Conference, ICVS 2019, Thessaloniki, Greece, Proceedings*, pages 723–733. Springer, 2019. 8

[83] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. 8

[84] Zhiyu Mou, Feifei Gao, Jun Liu, Xiang Yun, and Qihui Wu. Cluster head detection for hierarchical uav swarm with graph self-supervised learning. *IEEE Transactions on Signal Processing*, 72:5517–5532, 2024. 8

[85] Akarsh K Nair, Jayakrushna Sahoo, and Ebin Deni Raj. A lightweight fl-based uav detection model using thermal images. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security*, New York, NY, USA, 2024. Association for Computing Machinery. 5

[86] Ibrahim Nemer, Tarek Sheltami, Irfan Ahmad, Ansar Ul-Haque Yasar, and Mohammad A. R. Abdeen. Rf-based uav detection and identification using hierarchical learning approach. *Sensors*, 21(6), 2021. 3, 4

[87] Phuc Nguyen, Hoang Truong, Mahesh Ravindranathan, Anh Nguyen, Richard Han, and Tam Vu. Matthan: Drone presence detection by identifying physical signatures in the drone's rf communication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, page 211–224, New York, NY, USA, 2017. Association for Computing Machinery. 3

[88] University of Southern California Media Communications Lab (MCL). Usc drone detection and tracking dataset webpage. `http://mcl.usc.edu/mcl-drone-dataset/`, 2019. Accessed: Mar. 18, 2025. 2

[89] Anti-UAV Challenge Organizers. The 4th anti-uav workshop and challenge (cvpr 2025). `https://anti-uav.github.io/`, 2023. Accessed: Mar. 18, 2025. 2, 3

[90] UG2+ Challenge Organizers. The 7th ug2+ challenge (cvpr 2024). `https://ug2-uav-tracking.github.io/index.html`, 2024. Accessed: Mar. 22, 2025. 7

[91] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 8

[92] Seongha Park, Sangmi Shin, Yongho Kim, Eric T. Matson, Kiseon Lee, Paul J. Kolodzy, J. C. Slater, Matthew Scherreik, M. Sam, J. C. Gallagher, B. R. Fox, and Michael Hopmeier. Combination of radar and audio sensors for identification of rotor-type unmanned aerial vehicles (UAVs). In *Proceedings of the 2015 IEEE Sensors*, pages 1–4, Busan, South Korea, 2015. IEEE. 8

[93] Maciej Ł. Pawełczyk and Marek Wojtyra. Real world object detection dataset for quadcopter unmanned aerial vehicle detection. *IEEE Access*, 8:174394–174409, 2020. 6

[94] Duo Peng, Qiuhong Ke, ArulMurugan Ambikapathi, Yasin Yazici, Yinjie Lei, and Jun Liu. Unsupervised domain adaptation via domain-adaptive diffusion. *IEEE Transactions on Image Processing*, 2024. 7, 8

[95] Li-Hong Qiao, Rongxuan Wang, Yucheng Shu, Baobin Li, Weisheng Li, Xinbo Gao, and Zhanchuan Cai. Fast sampling of diffusion models for accelerated mri using dual manifold constraints. *IEEE Transactions on Circuits and Systems for Video Technology*, PP:1–1, 2025. 7, 8

[96] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 7, 8

[97] Aditya Ramesh, Prafulla Dhariwal, Alexander Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 7, 8

[98] Sairam VC Rebbapragada, Pranoy Panda, and Vineeth N Balasubramanian. C2fdrone: Coarse-to-fine drone-to-drone detection using vision transformer networks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6627–6633, 2024. 4, 6

[99] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 5

[100] Alejandro Rodriguez-Ramos, Javier Rodriguez-Vazquez, Carlos Sampedro, and Pascual Campoy. Adaptive inattentional framework for video object detection with reward-conditional training. *IEEE Access*, 8:124451–124466, 2020. 2, 4

[101] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 8

[102] Jakub Rosner, Tomasz Krzeszowski, Adam Świtoński, Henryk Josiński, Wojciech Lindenheim-Locher, Michał Zieliński, Grzegorz Paleta, Marcin Paszkuta, and Konrad Wojciechowski. Multimodal dataset for indoor 3d drone tracking. *Scientific Data*, 12(1):257, 2025. 2, 3

[103] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Detecting flying objects using a single moving camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):879–892, 2017. 2, 4

[104] Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real single-image flight without a single real image. In *Robotics: Science and Systems(RSS)*, 2017. 8

[105] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 7, 8

[106] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. 7, 8

[107] Stamatios Samaras, Eleni Diamantidou, Dimitrios Ataloglou, Nikos Sakellariou, Anastasios Vafeiadis, Vasilis Magoulianitis, Antonios Lalas, Anastasios Dimou, Dimitrios Zarpalas, Konstantinos Votis, Petros Daras, and Dimitrios Tzovaras. Deep learning on multi sensor data for counter uav applications—a systematic review. *Sensors*, 19 (22), 2019. 8

[108] Tushar Sangam, Ishan Rajendrakumar Dave, Waqas Sultani, and Mubarak Shah. Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6006–6013, 2023. 4

[109] Stirling Scholes, Alice Ruget, Germán Mora-Martín, Feng Zhu, Istvan Gyongy, and Jonathan Leach. Dronesense: The identification, segmentation, and orientation detection of drones via neural networks. *IEEE Access*, 10:38154–38164, 2022. 4, 5

[110] Ulzhalgas Seidaliyeva, Daryn Akhmetov, Lyazzat Ilipbayeva, and Eric T. Matson. Real-time and accurate drone detection in a video with a static background. *Sensors*, 20 (14), 2020. 4

[111] Hao Shen, Defu Lin, and Tao Song. A real-time siamese tracker deployed on uavs. *J. Real-Time Image Process.*, 19 (2):463–473, 2022. 7

[112] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 8

[113] Pranita Singh, Keshav Gupta, Amit Kumar Jain, Vishakha, Abhishek Jain, and Arpit Jain. Vision-based uav detection in complex backgrounds and rainy conditions. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pages 1097–1102, 2024. 6

[114] Chon Hou Sio, Yu-Jen Ma, Hong-Han Shuai, Jun-Cheng Chen, and Wen-Huang Cheng. S2siamfc: Self-supervised fully convolutional siamese network for visual tracking. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 1948–1957, New York, NY, USA, 2020. Association for Computing Machinery. 8

[115] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 2256–2265, 2015. 8

[116] Yang Song and Stefano Ermon. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 7, 8

[117] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020. 8

[118] Shizun Sun, Bo Mo, Junwei Xu, Dawei Li, Jie Zhao, and Shuo Han. Multi-yolov8: An infrared moving small object detection model based on yolov8 for air vehicle. *Neurocomput.*, 588(C), 2024. 4, 5

[119] Yumeng Sun, Jinguang Li, Linwei Wang, Junjie Xv, and yu Liu. Deep learning-based drone acoustic event detection system for microphone arrays. *Multimedia Tools and Applications*, 83, 2023. 5

[120] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. A dataset for multi-sensor drone detection. *Data in Brief*, 39:107521, 2021. 2, 5, 6, 8

[121] Fredrik Svanström, Cristofer Englund, and Fernando Alonso-Fernandez. Halmstad drone dataset (multi-sensor drone detection dataset) repository. https://github.com/DroneDetectionThesis/Drone-detection-dataset, 2021. GitHub repository, Accessed: Mar. 18, 2025. 2, 3

[122] Fredrik Svanström, Cristofer Englund, and Fernando Alonso-Fernandez. Real-time drone detection and tracking with visible, thermal and acoustic sensors. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7265–7272, 2021. 2, 4

[123] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, 2019. 7

[124] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 6

[125] Zhangyong Tang, Tianyang Xu, Zhenhua Feng, Xuefeng Zhu, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. Revisiting rgbt tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution, 2025. 7

[126] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. 8

[127] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[128] Xiaohan Tu, Chuanhao Zhang, Haiyan Zhuang, Siping Liu, and Renfa Li. Fast drone detection with optimized feature capture and modeling algorithms. *IEEE Access*, 12: 108374–108388, 2024. 4

[129] Zahoor Uddin, Aamir Qamar, Abdullah G. Alharbi, Farooq Alam Orakzai, and Ayaz Ahmad. Detection of multiple drones in a time-varying scenario using acoustic signals. *Sustainability*, 14(7), 2022. 5, 6

[130] Eren Unlu, Emmanuel Zenou, Nicolas Riviere, and Paul-Edouard Dupouy. Deep learning-based strategies for the detection and tracking of drones using several cameras. *IPSJ Transactions on Computer Vision and Applications*, 11(7), 2019. 5

[131] Viktor Walter, Matouš Vrba, and Martin Saska. On training datasets for machine learning-based visual relative localization of micro-scale uavs. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10674–10680, 2020. 3

[132] Bingshu Wang, Qiang Li, Qianchen Mao, Jinbao Wang, C. L. Philip Chen, Aihong Shangguan, and Haosu Zhang. A survey on vision-based anti unmanned aerial vehicles methods. *Drones*, 8(9), 2024. 5, 6

[133] Chuanyun Wang, Yang Su, Jingjing Wang, Tian Wang, and Qian Gao. Uavswarm dataset: An unmanned aerial vehicle swarm dataset for multiple object tracking. *Remote Sensing*, 14(11), 2022. 2, 3, 6, 7

[134] Chuanyun Wang, Linlin Meng, Qian Gao, Tian Wang, Jingjing Wang, and Linlin Wang. A target sensing and visual tracking method for countering unmanned aerial vehicle swarm. *IEEE Sensors Journal*, 24(19):30340–30351, 2024. 6, 7

[135] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023. 5, 6

[136] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, pages 1–21. Springer, 2024. 7

[137] Ning Wang, Wengang Zhou, Yibing Song, Chao Ma, Wei Liu, and Huchuan Li. Unsupervised deep representation learning for real-time tracking. *International Journal of Computer Vision*, 129(2):400–418, 2021. 8

[138] Ye Wang, Yueru Chen, Jongmoo Choi, and C.-C. Jay Kuo. Towards visible and thermal drone monitoring with convolutional neural networks. *APSIPA Transactions on Signal and Information Processing*, 8:e5, 2019. 2

[139] Mariusz Wisniewski, Zeeshan A. Rana, and Ivan Petrunin. Drone model classification using convolutional neural network trained on synthetic data. *Journal of Imaging*, 8(8), 2022. 3, 5

[140] Mariusz Wisniewski, Zeeshan A. Rana, and Ivan Petrunin. Synthetic drone classification dataset, 2022. 2, 3, 5

[141] Mariusz Wisniewski, Zeeshan A. Rana, Ivan Petrunin, Alan Holt, and Stephen Harman. Drone detection using deep neural networks trained on pure synthetic data, 2024. 4

[142] Mariusz Wisniewski, Zeeshan A. Rana, Ivan Petrunin, Alan Holt, and Stephen Harman. Towards fully autonomous drone tracking by a reinforcement learning agent controlling a pan–tilt–zoom camera. *Drones*, 8(6), 2024. 8

[143] Guangyu Wu, Fuhui Zhou, Kai Kit Wong, and Xiang-Yang Li. A vehicle-mounted radar-vision system for precisely positioning clustering uavs. *IEEE Journal on Selected Areas in Communications*, 42(10):2688–2703, 2024. 4, 5, 6, 8

[144] Tongyan Wu, Haibin Duan, and Zhigang Zeng. Biological eagle-eye-based correlation filter learning for fast uav tracking. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024. 6, 7

[145] Xin Wu, Wei Li, Danfeng Hong, Ran Tao, and Qian Du. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):91–124, 2022. 7

[146] Dingkun Xiao, Zhenzhong Wei, and Guangjun Zhang. Target-aware transformer tracking with hard occlusion instance generation. *Frontiers in Neurorobotics*, 17, 2024. 7

[147] Zhenyuan Xiao, Yizhuo Yang, Guili Xu, Xianglong Zeng, and Shenghai Yuan. Av-dtec: Self-supervised audio-visual fusion for drone trajectory estimation and classification, 2024. 3, 5

[148] Zhenyuan Xiao, Huanran Hu, Guili Xu, and Junwei He. Tame: Temporal audio-based mamba for enhanced drone trajectory estimation and classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 3, 5

[149] Xueli Xie, Jianxiang Xi, Xiaogang Yang, Ruitao Lu, and Wenxin Xia. Stftrack: Spatio-temporal-focused siamese network for infrared uav tracking. *Drones*, 7(5), 2023. 6

[150] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457, 2021. 7

[151] Yizhuo Yang, Shenghai Yuan, Jianfei Yang, Thien Hoang Nguyen, Muqing Cao, Thien-Minh Nguyen, Han Wang, and Lihua Xie. Av-fdti: Audio-visual fusion for drone threat identification. *Journal of Automation and Intelligence*, 3(3):144–151, 2024. 8

[152] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8896–8905, 2022. 7, 8

[153] Xinyi Ying, Chao Xiao, Wei An, Ruojing Li, Xu He, Boyang Li, Xu Cao, Zhaoxu Li, Yingqian Wang, Mingyuan Hu, Qingyu Xu, Zaiping Lin, Miao Li, Shilin Zhou, Weidong Sheng, and Li Liu. Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2025. 1, 2

[154] Ryota Yoshihashi, Tu Tuan Trinh, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Differentiating objects by motion: Joint detection and tracking of small flying objects, 2018. 4

[155] Qianjin Yu, Yinchao Ma, Jianfeng He, Dawei Yang, and Tianzhu Zhang. A unified transformer-based tracker for anti-uav tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3036–3046, 2023. 6, 7

[156] Shenghai Yuan, Yizhuo Yang, Thien Hoang Nguyen, Thien-Minh Nguyen, Jianfei Yang, Fen Liu, Jianping Li, Han Wang, and Lihua Xie. Mmaud: A comprehensive multi-modal anti-uav dataset for modern miniature drone threats. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2745–2751, 2024. 1, 2, 3, 5, 6, 7

[157] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021. 7, 8

[158] Junjie Zhang, Yi Lin, Xin Zhou, Pangrong Shi, Xiaoqiang Zhu, and Dan Zeng. Precision in pursuit: A multi-consistency joint approach for infrared anti-uav tracking. *The Visual Computer*, 41:2187–2202, 2024. 6

[159] 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 *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland. 7

[160] Yulong Zhang, Shuhao Chen, Weisen Jiang, Yu Zhang, Jiangang Lu, and James T. Kwok. Domain-guided conditional diffusion model for unsupervised domain adaptation. *Neural Networks*, 184:107031, 2025. 7, 8

[161] Zhihao Zhang, Lei Jin, Shengjie Li, JianQiang Xia, Jun Wang, Zun Li, Zheng Zhu, Wenhan Yang, PengFei Zhang, Jian Zhao, and Bo Zhang. Modality meets long-term tracker: A siamese dual fusion framework for tracking uav. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1975–1979, 2023. 6, 7

[162] Jian Zhao, Gang Wang, Jianan Li, Lei Jin, Nana Fan, Min Wang, Xiaojuan Wang, Ting Yong, Yafeng Deng, Yandong Guo, Shiming Ge, and Guodong Guo. The 2nd anti-uav workshop & challenge: Methods and results, 2021. 2, 3, 6, 7

[163] Jie Zhao, Jingshu Zhang, Dongdong Li, and Dong Wang. Vision-based anti-uav detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 25323–25334, 2022. 1, 2, 4

[164] Jian Zhao, Jianan Li, Lei Jin, Jiaming Chu, Zhihao Zhang, Jun Wang, Jiangqiang Xia, Kai Wang, Yang Liu, Sadaf Gulshad, Jiaojiao Zhao, Tianyang Xu, Xuefeng Zhu, Shihan Liu, Zheng Zhu, Guibo Zhu, Zechao Li, Zheng Wang, Baigui Sun, Yandong Guo, Shin ichi Satoh, Junliang Xing, and Jane Shen Shengmei. The 3rd anti-uav workshop & challenge: Methods and results, 2023. 2, 3, 6, 7

[165] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham. 3d motion capture of an unmodified drone with single-chip millimeter wave radar. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5186–5192, 2021. 2, 4, 5, 6

[166] Ye Zheng, Zhang Chen, Dailin Lv, Zhixing Li, Zhenzhong Lan, and Shiyu Zhao. Air-to-air visual detection of micro-uavs: An experimental evaluation of deep learning. *IEEE Robotics and Automation Letters*, 6(2):1020–1027, 2021. 2, 3, 4, 5

[167] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022. 7, 8

[168] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision – ECCV 2022*, pages 350–368, Cham, 2022. Springer Nature Switzerland. 8

[169] Yue Zhou, Yutong Jiang, Zhonglin Yang, Xingxin Li, Wenchen Sun, Han Zhen, and Ying Wang. Uav image detection based on multi-scale spatial attention mechanism with hybrid dilated convolution. In *2024 3rd International Conference on Image Processing and Media Computing (ICIPMC)*, pages 279–284, 2024. 4

[170] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge, 2018. 5, 6

[171] Xue-Feng Zhu, Tianyang Xu, Jian Zhao, Jia-Wei Liu, Kai Wang, Gang Wang, Jianan Li, Qiang Wang, Lei Jin, Zheng Zhu, Junliang Xing, and Xiao-Jun Wu. Evidential detection and tracking collaboration: New problem, benchmark and algorithm for robust anti-uav system, 2023. 2, 8

[172] Ender Çetin, Cristina Barrado, and Enric Pastor. Counter a drone in a complex neighborhood area by deep reinforcement learning. *Sensors*, 20(8), 2020. 8

[173] Ender Çetin, Cristina Barrado, and Enric Pastor. Countering a drone in a 3d space: Analyzing deep reinforcement learning methods. *Sensors*, 22(22), 2022. 8