A Comprehensive Review for Typical Applications Based Upon Unmanned Aerial Vehicle Platform

Yuqi Han, Huaping Liu, *Senior Member, IEEE*, Yufeng Wang, *Member, IEEE*, and Chunlei Liu

***Abstract*—Unmanned aerial vehicles (UAVs) have been widely applied in military and civilian fields due to their flexibility and effectiveness. As a vital component of UAVs, the vision system has taken on great significance in different applications (e.g., autonomous landing, traffic surveillance, and disaster rescue) to attract widespread attention in recent years. Therefore, the automatic understanding of visual data collected from these air platforms becomes urgently needed in UAV systems. In this review, we revisit and summarize the recent techniques and developments for several typical UAV applications, including object detection, object tracking, and semantic segmentation. In addition, we also highlight the difficulties and subsequent orientations from different perspectives, which may stimulate future research and applications in the UAV vision era.**

***Index Terms*—Object detection, object tracking, semantic segmentation, unmanned aerial vehicles (UAV).**

I. Introduction

As THE remote sensing platforms and technologies have been leaping forward [1] over the past few years, how to effectively and intelligently process and interpret massive data has attracted widespread attention among researchers. As one of the critical remote sensing platforms, unmanned aerial vehicles (UAVs) can obtain wide-view vision and multimodal information in real time for the postprocessing in different regimes [2], [3], [4], [5], [6]. However, different from the general monitoring scenes, where the cameras (e.g., mobile phones, handheld cameras, surveillance cameras, and satellite) are always static on the ground or slow-moving with less geometric and photo-metric changes [7], [8], [9], [10], [11]. The UAV platforms overlook the targets from the air, which enables it to acquire the data flexibly and make up the information loss for the target appearance due to the geographic and time limitations. However, its unique imaging mechanism and characteristic also pose new challenges for remote sensing vision tasks.

Inevitable Image Degradation: Considering the rapid movement for the UAV platform and the target of the interest, the external environment changes rapidly (e.g., weather, illumination condition, scenes). Moreover, under strong wind, the platform would usually inevitably undergo mechanical vibration, which may even result in motion blur and fuzzy image degradation. Such challenging attributes may bring in a large variety of object appearance, which degrades the quality for the captured data. In addition, harsh scenes, such as rainy or foggy days and night, which has poor visibility, also bring new challenges for the algorithms to detect the object from the background. Therefore, to improve the quality of the captured data, it would be necessary to carry out a preprocessing image module to reduce the noise and correct the camera distortion.

Uneven Target Size and Distribution: Generally, the UAV obtains data from different altitudes using a large aperture, fixed focal, and wide-angle lens, thus resulting in an uneven target size problem. Specifically, some objects may be densely located, even overlap with each other, while some objects may be very sparse; some objects may occupy a large proportion of the image, while some objects are very small with limited distinct features. In [12], Han pointed put that such uneven statistical properties would also increase the difficulties of detecting the targets from their surrounding background.

Viewpoint Variation and Occlusion: Due to the UAV platforms having the characteristics of large freedom and mobility degrees, UAVs might capture the targets from different aspects by flying around the targets by 360°. For example, UAVs can capture the back or front side of the targets, in which case the targets may have severe variations in the imaging process [13]. This will become a big challenge if the methods do not have the ability for timely online learning and model updates. In addition, partial or even full occlusion is common due to the high mobility freedom of the UAV platform, as illustrated in [14]. However, such attributes would temporally corrupt the target template and may lead to detection failure and tracking drift due to model degradation.

Limited Computation Source: For most of the UAV platforms, only a single CPU could be embedded as the processing resources due to the strict limitations in terms of its weight and power, which greatly limits the on-board computing speed. To this end, the intelligent algorithms should be carefully designed without casting aside high efficiency in order to meet the

Manuscript received 29 June 2022; revised 3 September 2022 and 30 September 2022; accepted 14 October 2022. Date of publication 25 October 2022; date of current version 14 November 2022. This work was supported by the China Postdoctoral Science Foundation under Grant 2021TQ0177. *(Corresponding author: Chunlei Liu.)*

Yuqi Han and Huaping Liu are with the Beijing National Research Center for Information Science and Technology, Institute for Artificial Intelligence, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yuqi_han@tsinghua.edu.cn; hpliu@tsinghua.edu.cn).

Yufeng Wang is with the Institute of Unmanned System, Beihang University, Beijing 100191, China (e-mail: wyfeng@buaa.edu.cn).

Chunlei Liu is with the School of Electrical and Information Engineering, Beihang University, Beijing 100191, China (e-mail: liuchunlei19940430@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3216564

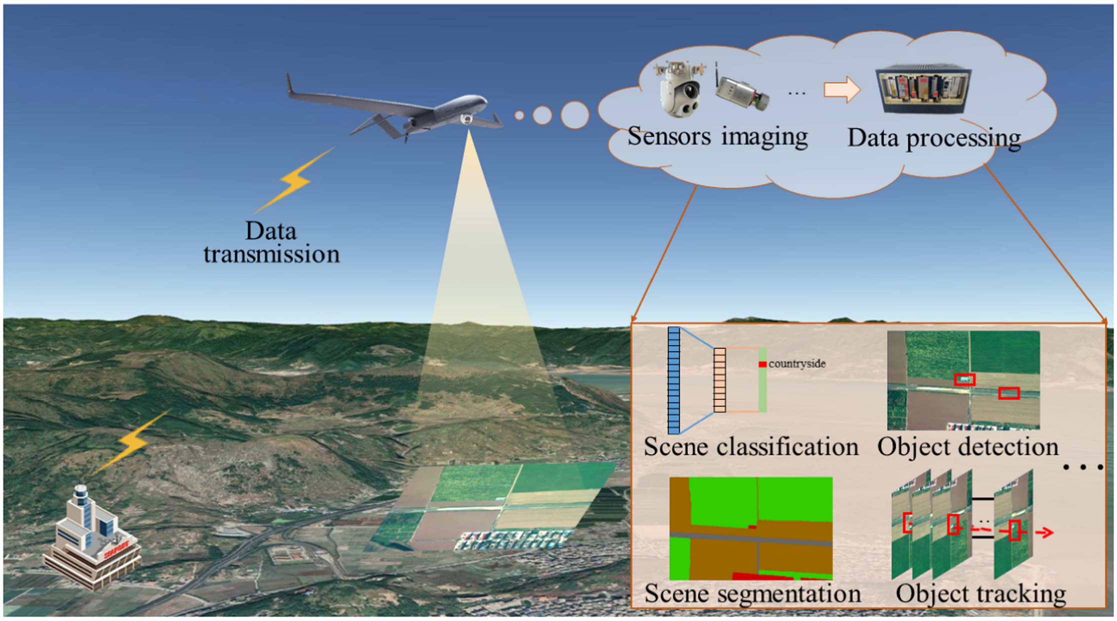

Fig. 1. Schematic diagram for several typical UAV-based Earth Observation applications. The UAVs would capture and pre-process the data with the on-board sensors firstly. Afterward, they would transmit the captured data too the ground station. The ground station would analyze the obtained data with according to the category for different applications.

real-time requirement for on-board processing. In addition, considering of the energy-consuming applications like maneuvering flight, the on-board algorithms also need to be light-weighted enough to save the power supplies at best.

These issues present significant challenges in analyzing the image and video data captured from the UAV platform. Aiming for these challenges, many works have emerged to extract useful information from the data, and perform different tasks of UAVs, thus making the UAV more intelligent.

Notably, researchers have concentrated on various UAV-based vision tasks with cutting-edge deep learning (DL) technologies. Some studies summarize the current research on one specific task of UAVs. However, most reviews target normal camera objects [15], [16], [17], [18], [19], [20], while few reviews focus on UAV-view objects [21], [22], [23]. To the best of our knowledge, there is a lack of survey on UAV emergency landing. Therefore, based on the practical background of the Earth observation, we provide a unified overview of the object detection, tracking, and semantic segmentation technologies of images and videos captured from the UAVs. The main UAV-based Earth observation scene can be seen in Fig. 1. First, UAVs acquire data and preprocess the data using corresponding sensors; second, UAVs transmit the data back to the ground station, and the ground station performs task analysis, including but not limited to, scene classification, target detection, target segmentation, scene segmentation, etc. Our work highlights the following aspects.

1) Unlike other works that only review single tasks or object detection/tracking tasks, our work aims for the typical applications, i.e., object detection, object tracking, and sematic segmentation for UAV Earth observation. It should be noted that object tracking indicates the single object tracking.

2) This article focuses on analyzing various representative and recent algorithms thoroughly. We found that the existing methods are mostly evaluated on a specific dataset, and a comprehensive benchmark is lacking.

3) We summarize the challenges in the UAV imaging process and provide future directions, which could benefit the audience in the UAV vision area.

The overall structure of this study is organized as follows. In Section II, we present a brief description of UAV-based datasets for the abovementioned applications. Section III provides a detailed description of the relevant works and algorithms for these applications. In Section IV, we discuss the potential directions to stimulate the development of this field. Finally, Section V concludes this article.

II. Dataset

Noting that there exist numerous aerial images and video datasets for object detection, single object tracking and semantic segmentation (e.g., DOTA [49], NWPU VHR-10 [50], and VEDAI [51]). In this study, we will focus on reviewing the datasets captured on the UAV platform. Illustration and featured attributes, including the length or sequences, total representative frames, target categories, and their corresponding available websites are shown in Tables I and II.

A. Object Detection

Okutama-Action: Okutama-Action [24] is a human action detection dataset captured from at 45/90° cameras mounted at two flexible UAV platforms in 2017. It is formed with 43 fully annotated sequences containing 12 actions, including carrying handshaking, drinking, and reading, with 77 365 total frames.

TABLE I
Illustration and Featured Attributes for the Existing UAV-Captured Datasets in Typical Earth-Observation Applications

Datasets	Year	Sequences/length	Total Frames/images/targets	Categories	Applications
Okutama-Action [24]	2017	43	77 365	12	detection
VisDrone2022-DET [25]	2022	-	10 209	10	detection
VisDrone2021-DET [25]	2021	-	8599	10	detection
VisDrone2020-DET [25]	2020	-	8599	10	detection
VisDrone2019-DET [25]	2019	-	8599	10	detection
VisDrone2018-DET [25]	2018	-	8599	10	detection
MOR-UAV [26]	2020	30	10 948	2	detection
CARPK [27]	2017	-	1448	1	detection
AU-AIR [28]	2020	-	32 823	8	detection
UVSD [29]	2020	-	5874	2	detection
DroneVehicle [30]	2021	-	441 642	5	detection
BIRDSAI [31]	2020	172	-	-	detection
MOHR [32]	2021	-	90 014	5	detection
VSAI [33]	2022	-	49 712	2	detection
UAVDT [34]	2018	100	∼80 000	3	detection; tracking
Stanford Drone Dataset [35]	2016	-	929 499	6	detection
HighD [36]	2018	16.5h	110 000	1	detection; tracking
DTB70 [37]	2017	70	15 777	-	tracking
VisDrone2020 [25]	2020	192	221 920	10	tracking
VisDrone2019 [25]	2019	167	188 998	10	tracking
VisDrone2018 [25]	2018	132	106 354	10	tracking
UAV123 [38]	2016	123	>110 000	-	tracking
UAV20L [38]	2016	20	58 670	-	tracking
Anti-UAV [39]	2021	736	5 859 000	-	tracking
Small90 [40]	2019	90	∼39 380	-	tracking
Small112 [40]	2019	112	∼55 669	-	tracking
UAVDark135 [41]	2022	135	125 466		tracking
DarkTrack2021 [42]	2022	110	100 000	-	tracking
UAVTrack112 [43], [44]	2021	112	100 313	-	tracking
AVSD [45]	2019	10	525	6	segmentation
UAVid [46]	2018	30	300	8	segmentation
AeroScapes [47]	2018	-	3269	11	segmentation
ManipalUAVid [48]	2019	-	667	4	segmentation

The recording UAV works at the height of 10–45 m, with a 30-fps imaging speed and 3840 × 2160 image resolution. Okutama-Action dataset gathers several typical challenging factors in the action detection field, such as abrupt camera movement, remarkable aspect ratio and scale variation, and dynamic action transition.

VisDrone: The VisDrone dataset [25] is collected by the AISKYYE team of the Machine Learning and Data Mining Laboratory of Tianjin University. The image data in this dataset are collected by different types of drones from 14 cities (both country and urban) across China with a variety of lightening and weather conditions (i.e., daytime, night, rainy, and foggy). The initial construction for this dataset starts in 2018. Afterward, several object detection and tracking challenges are host in top-ranking computer vision conferences from then on. In addition, the capacity and difficulty of the dataset would continue to increase over the previous year. More specifically, in 2018, the organizing committee provided 8599 representative frames with ten classes of targets (i.e., pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle). While in 2022, the dataset supplies 400 videos, including 10 209 static pictures and 265 228 frames to fully validate the performance for participated algorithms. It should be also noted that in VisDrone, most of the targets are densely distributed or overlapped in 2.6 million labeled bounding boxes. Besides that, some targets are extremely small, which also pose great challenges for detectors to generate suitable anchors. Some crucial factors, such as out of view and occlusion factors are also highlighted in the ground-truth for better and more accurate validation.

MOR-UAV: MOR-UAV [26] dataset is collected and made public by the research team in Malaviya National Institute of Technology Jaipur in 2020. This dataset contains 30

TABLE II
ILLUSTRATION FOR THE AVAILABLE WEBSITE SOURCES OF THE REVIEWED UAV-BASED DATASETS

Datasets	Website Source
Okutama-Action [24]	http://okutama-action.org
VisDrone2018-2022 [25]	https://github.com/VisDrone/VisDrone-Dataset
MOR-UAV [26]	https://visionintelligence.github.io/Datasets.html
CARPK [27]	https://paperswithcode.com/dataset/carpk
AU-AIR [28]	https://bozcani.github.io/auairdataset
DroneVehicle [30]	https://github.com/VisDrone/DroneVehicle
UVSD [29]	https://github.com/liuchunsense/UVSD
UAVDT [34]	https://sites.google.com/view/grli-uavdt/
BIRDSAI [31]	https://sites.google.com/view/elizabethbondi/dataset
Stanford Drone Dataset [35]	https://cvgl.stanford.edu/projects/uav_data/
HighD [36]	https://www.highd-dataset.com
DTB70 [37]	https://github.com/flyers/drone-tracking
UAV123/20L [38]	https://cemse.kaust.edu.sa/ivul/uav123
Anti-UAV [39]	https://github.com/ucas-vg/Anti-UAV
Small90/112 [40]	https://github.com/bczhangbczhang/smallobject
UAVDark135 [41]	https://vision4robotics.github.io/project/uavdark135/
DarkTrack2021 [42]	https://darktrack2021.netlify.app
UAVTrack112 [43], [44]	https://github.com/vision4robotics/SiamAPN
AVSD [45]	https://github.com/wyfeng1020/AVSD
UAVid [46]	https://uavid.nl
AeroScapes [47]	https://github.com/ishann/aeroscapes
ManipalUAVid [48]	https://github.com/uverma/ManipalUAVid

UAV-captured video sequences with 10 948 frames at different locations (e.g., highways, agricultural regions, urban areas, and traffic intersections) with various of challenging attributes (flexible viewpoint, altitude, abrupt drone motion, changing lightning conditions, different weather, occlusion, and temporal out-of-view). The authors have categorized the captured 89 783 instances into two classes, i.e., cars and heavy vehicles. The sequences are recorded at 30 fps with the image resolution varying from 1280 × 720 to 1920 × 1080 in MOR-UAV. In addition, the moving instances are automatically labeled using YOLO-mark tool, which would be employed for validating target detection and recognition algorithms.

Stanford Drone: Stanford Drone [52] is a large-scale object detection and tracking dataset, which was collected by Stanford University in 2016. This dataset mainly collects the outdoor scenarios for the Stanford university campus with a 4 K camera mounted at a quad-copter at the height of approximate 80 m above the ground. Afterward, the collected videos are processed and generates a series of image sequences with the resolution of 1400 × 1904. In sum, 929 499 frames with six categories are carefully labeled in the Stanford Drone. Specifically, this dataset cover over 19 000 targets, including 11 000 buses, 22 000 golf carts, 33 000 skateboarders, 13 000 cars, 64 000 bicyclists, and 112 000 walking pedestrians.

UAVDT: UAVDT [34] was collected by the Chinese Academy of Sciences in 2018, which aims to provide a unified large-scale benchmark for multiple tasks, such as vehicle tracking and detection. In UAVDT, 100 sequences from nearly 10 h of raw videos are selected and processed into about 80 000 annotated representative frames at various common scenes, including toll stations, highways, arterial roads, intersections, squares, and so on. About 2700 vehicles are broadly categorized into three classes (i.e., car, truck, and bus), with 840 000 annotated bounding boxes. Furthermore, the image resolution in UAVDT dataset is 1080 × 540 pixels and the imaging speed is 30 fps. Comparing to the other existing datasets, UAVDT contains up to 14 representative challenging factors in detection and tracking (i.e., occlusion, vehicle category, camera view, flying altitude, and weather condition).

CARPK: The Car Parking Lot dataset (CARPK) [27] is proposed by the National Taiwan University, in 2017, which collects 1448 images with approximately 90 000 cars in accordance at four different parking lots. Apart from vehicle detection, CARPK is also the first large-scale UAV-captured dataset to validate counting algorithms, where each vehicle target is manually annotated for facilitating evaluation. As for the details for this CARPK, the image sequences are collected by a Phantom 3 drone with the flying altitude at about 40-m high. In CARPK, the largest target is much more bigger than 64 × 64, and the maximum number of the targets in a single view is about 200, which reflects the characteristic for the target (multiscale and dense distributed) in parking counting application.

AU-AIR: AU-AIR [28] is a multimodal UAV object detection dataset, organized by the Department of Engineering, Aarhus University, in 2020. Different from the other UAV-collected dataset, AU-AIR not only provides the visual data but also supplies the other necessary modal information (i.e., the current altitude, velocity, GPS, time, and IMU). It has 32 823 labeled frames, at the size of 1920 × 1080 pixels. AU-AIR collects 132 034 object instances with eight object categories (namely, pedestrian, trailer, bike, motorbike, car, bus, truck, and van) related to traffic surveillance in various of weather and lightening circumstances.

UVSD: UVSD [29] is collected and formed by Shandong University with a DJI matrice 200 platform in a variety of locations and altitudes in 2020. This dataset is made up of 5874 images with their resolution varies from 5280 × 2970 to 960 × 540 pixels. In addition, the vehicle instances in UVSD are densely distributed with more than 150 vehicles per image. To this end, UVSD could also be employed for the validation for other task, such as vehicle counting. Furthermore, UVSD contains up to 98 600 high-quality annotations with different types, including horizontal bounding-box, oriented bounding-box, as well as instance-level semantic annotations.

DroneVehicle: In order to overcome the low light conditions in UAV visual tasks, DroneVehicle dataset [30] collects 15 532 RGB-Thermal image pairs and 441 642 instances with their resolution at 840 × 712 pixels. The dataset mainly focuses on the urban field covering roads, parking lots, residential fields, highways, and so on. In order to validate the performance for vehicle detection and counting approaches, the UAV platform collects the image sequences from day to night, with large-scale illumination variation. The authors divide the vehicles into five categories, but the number for each class is somehow unbalanced according to the statistical counting in the original article.

BIRDSAI: BIRDSAI [31] (pronounced similar to bird's eye) is a large-scale infrared UAV object detection and tracking dataset, organized by the Harvard University in 2020. Similar to UAV123, BIRDSAI contains both the real aerial videos and synthetic sequences. Specifically, 48 real infrared sequences are collected with changing wavelength on a fixed-wing UAV in multiple protected areas at southern Africa, while 124 synthetic sequences are generated from Air Shepherd. The image frame in BIRDSAI contains lots of challenging factors which may affect stable tracking and accurate detection, such as image rotation, target deformation, large-scale change, and aspect ratio variation, etc. The resolution for each respective frame is fixed at 640 × 480 pixels.

MOHR: MOHR dataset [32] is a TIR object detection benchmark, collected by Harbin Institute of Technology in 2020 to extend the object detection research for large-scale variation, arbitrary orientations, as well as irregular target deformation. In MOHR, 90 014 object instances could be broadly classified into five categories, namely, building, flood damage, truck, car, and collapse. Furthermore, in order to quantitatively validate the performance for the testing detectors, the authors manually annotate this dataset and count the number for each class as follow. In MOHR, there are 41 468 buildings, 25 575 cars, 12 957 trucks, 7718 flood damages, as well as 2296 collapses, with a large-range of scale changes. It should be noted that collapses and flood damages are first concluded as target categories in UAV dataset. Furthermore, MOHR is collected with three types of cameras (Nikon D800, Sonny RX1rM2, as well as DJI Phantom 4Pro) at varying flying height. In this way, 3048 aerial images have the size of 5482 × 3078, 5192 images have the size of 7360 × 4912. While for the rest of 2390 screenshots, their resolution is 8688 × 5792.

VSAI: VSAI dataset [33] is a dataset for object detection, which was collected by the National University of Defense Technology in 2022. In VSAI, 444 images are collected by different camera angles, flight height, times, weather conditions and illuminations. VSAI contains the bounding boxes of objects with two shapes, i.e., oriented bounding boxes (49 712) and arbitrary quadrilateral bounding boxes (47 519 small vehicles and 2193 large vehicles). The resolution of these data includes 4000 × 3000, 5472 × 3648, and 4056 × 3040. In order to further improve the generalization abilities of the object detection methods, VSAI also annotates the occlusion rate of objects.

B. Object Tracking

DTB70: DTB70 [37] comprises 70 video sequences with 15 777 representative frames. The construction for this dataset is led by the Hong Kong University of Science and Technology, in 2017. What is interesting, DTB70 is made up of two constitute parts, where some of the sequences collects the outdoor scenarios for the university campus with a 4 K camera mounted at a DJI Phantom-2 drone flying at the height of approximate 120-m high. While the other parts are supplemented from Youtube to introduce the diversity for the data distribution. Each frames are carefully annotated with horizontal bounding box same as some other UAV datasets, and the resolution for the respective video frame is 1280 × 720.

VisDrone: In addition to the object detection, VisDrone dataset also has the challenge sequences for object tracking. VisDrone 2018 single-object tracking task dataset contains 132 sequences with about 106 354 frames. Based on these data, VisDrone2019 provides 167 challenging sequences with 188 998 frames in total. Furthermore, VisDrone2020 provides 192 challenging sequences with 221 920 frames in total.

UAV123: UAV123 [38] is collected and proposed by KAUST (King Abdullah University of Science and Technology), in 2016, which involves 123 sequences and over 110 000 representative images from an aerial viewpoint. UAV123 is made up of three parts, including a professional DJI UAV, a tiny UAV with low cost, and a self-designed UAV simulator. Therefore, the resolution for the respective frame varies due to the difference for the captured platform. Aiming at supplement the gap for aerial object tracking, each frame is carefully annotated by the authors with horizontal bounding boxes and its corresponding attributes (i.e., occlusion, camera motion, illumination variation, aspect ratio change). In addition, the flying circumstance for these UAV platforms varies a lot (i.e., weather condition, flying altitude, scenarios), in order to enhance the variety and challenges for this dataset.

UAV20 L: The authors select 20 long-term video sequences in UAV123 to form up UAV20 L dataset. As a subset for UAV123, UAV20 L has 58 670 frames in total. According to the experimental results reported in the original article, most of the testing trackers perform inferior in UAV20 L when compared with their performance in UAV123. Such phenomenon could be attributed to the absence for redetection mechanism for the testing trackers, which has also pointed out a direction for object tracking field.

Anti-UAV: Anti-UAV [39] is collected by research team in the University of Chinese Academic of Sciences using two types of drones (DJI and Parrot) in 2021. The initial purpose for publishing Anti-UAV is to pioneer an interesting research field in the task of tracking UAV. Anti-UAV dataset comprises 318 RGB-T video pairs containing 585.9 k annotations, where the respective pair contains a thermal video and an RGB video. The videos cover a variety of backgrounds (e.g., tree, cloud, building), two light modes (visible and infrared), and two lighting conditions (night and day) at 25 FPS.

Small90/112: Small90 [40] comprises 90 small-sized object sequences with about 39 380 frames, in which additional challenges (e.g., low resolution and target drifting) are encompassed. Based on Small90, Small112 [35] adds another 20 more challenging sequences. Totally, Small112 has 112 sequences with about 55 669 frames.

UAVDark135: UAVDark135 [41] refers to the first dark tracking benchmark based upon UAV platform, to make up the blank for tracking performance evaluation in dark environment. UAVDark135 comprises 135 sequences filmed with a standard UAV at night with 125 466 frames with manual annotation. The total frames, mean frames, maximum frames, and minimum frames of the benchmark are 125 466, 929, 4571, and 216, respectively. Meanwhile, UAVDark135 contains various scenes (e.g., lakeside, highway, street, ocean, and road) and covers considerable objects (e.g., bikes, trucks, athletes, buildings, cars, and pedestrians) making it suitable for large-scale evaluation.

HighD: Researchers in Aachen University collect 16.5 h of measurements to form up HighD [36] dataset, which contains 110 000 vehicles with 5600 recorded lane variation and 45 000 km driving distance in total. The videos are collected using a consumer quad-copter at the recording rate of 25 fps with the image resolution set as 4096 × 2160. In addition, HighD involves six different recording locations with different traffic circumstances during sunny and windless weather from 8 A.M. to 5 P.M. Different from other datasets, HighD is organized initially for safety assessment, but it could also be employed into the simulation and validation for vehicle counting, traffic analysis, and object tracking.

DarkTrack2021: DarkTrack2021 [42] covers 110 challenging sequences with 100 K frames, which are taken with 30 FPS at night-time in urban scenes. Similar to UAVDark135 [41], the original purpose for constructing this dataset is to provide a comprehensive assessment for tracking performance in ill-lighting status. The shortest, longest, and the average length of sequences are 92, 6579, and 913 frames, respectively. DarkTrack2021 provides abundant scenarios of in night-time real world with various challenges, including full-occlusion, low resolution, motion blur, and viewpoint variation.

UAVTrack112: UAVTrack112 [43], [44] is created from images captured and annotated during the real-world tests, which contains 112 sequences with 100 313 representative frames. The aim of establishing this dataset is for aerial tracking. Therefore, some cityscape scenes are also selected in this dataset. Same as DarkTrack2021 [42], this dataset is organized and maintained by Tongji University, China.

C. Semantic Segmentation

AVSD: AVSD [45] is designated as public by Beihang University, in 2020, which involves ten different sequences with total 525 pictures. 131 pictures out of all the 525 pictures are annotated manually. The sequences are captured at the speed of 12 fps with their resolution fixed as 1280 × 1024. In addition, there are six classes of targets in AVSD, namely, bare land, grassland, forest, building, road, and vehicles. The most challenging factor for AVSD is the variant motion and scene complexity for the collected video sequences.

UAVid: UAVid dataset [46] is jointly constructed by University of Twente and Wuhan University, in 2018, including 30 video sequences with the image resolution fixed as 4 K. In this dataset, 300 pictures are densely labeled with eight classes (i.e., background clutters, moving cars, humans, low vegetation, trees, static cars, roads, and buildings) for the urban scene understanding task. Noting that the authors also propose an in-house video labeling tool to automatically annotate the sequences in UAVid.

AeroScapes: The AeroScapes aerial semantic segmentation benchmark [47], was designed and organized by Carnegie Mellon University, in 2018. The imaging altitude for the commercial drone varies from 5-m high to 50-m high when constructing this dataset. According to the original article, AerosScapes is made up of 3269 pictures for 11 object classes with large-scale variation, viewpoint change, as well as scenarios composition.

ManipalUAVid: ManipalUAVid [48] is constructed and made publicly available by Manipal Institute of Technology, in 2019, which comprises 667 frames with four classes: road, construction, greenery, and water bodies. They are captured in six locations, such as the library, canteen, and hostel, with an image resolution of 1280 × 720. The presentation for ManupalUAVid greatly complements the gap in the direction of semantic segmentation using UAV platform.

III. Methods

Taking the emergency landing of UAVs as the application background, this section presents a brief overview of methods for object detection, object tracking, and semantic segmentation of the UAV images and videos.

A. Object Detection

Recent advancements in deep learning technologies create large opportunities to study object detection in a previously inaccessible way. Existing object detection methods can usually be divided into two types: 1) two-stage detectors, where one model is adopted for the extraction of object region proposals and another model is adopted for classifying and refining the object localization, including fast R-CNN [53], faster R-CNN [54], cascade RCNN [55], etc. 2) One-stage detectors refer to models skipping the region proposal stage of the two-stage models and implementing detection over a dense sampling of locations, including the YOLOv1 [56], YOLOv2 [57], SSD [58], RetinaNet [59], FCOS [60], etc. In general, the two-stage detectors achieve higher object localization and recognition accuracy, while the one-stage detectors are characterized by higher inference speed. Next, we introduce the detecting methods for the UAV environment in detail. Mittal et al. [21] reviewed the low-altitude UAV object detection based on deep learning. They proposed that low-altitude UAV-based object detection has more challenges compared with standard images, such as large-scale changes, densely distributed objects, arbitrary orientations, object relative motion, detection for small objects, class imbalance, and large-scale changes. In the following, we mainly review the representative deep learning-based UAV object detection methods [61], [62], [63], [64] with detectors at one stage and two stages.

One-stage Detector: To mitigate the real-time scene parsing challenges, Zhang et al. [65] developed the SlimYOLOv3 model to be capable of learning efficient deep object detectors via channel pruning of convolutional layers. For the problem of small object detection, Liang et al. [66] proposed a feature fusion and scaling-based single shot detector (FS-SSD), which incorporates the spatial object relationships into object redetection. To tackle the small objects in UAV images, Liu et al. [67] proposed a multiscale feature fusion algorithm, termed as dilated-attention-feature fusion SSD (D-A-FS SSD), with the combination of dilated convolution and attention mechanism. Liu et al. [68] developed UAV-YOLO for detecting small objects in UAV by enlarging the receptive field. To tackle the challenges of large-scale change and real-time problems, Li et al. [69] proposed the DSYolov3 model by adding multiple scale-aware decision discrimination networks, which involves a channel attention model and a sparsity-based channel prunning based on the YOLOv3 model.

Two-/Multistage Detector: To increase the resolution of objects in UAV images, Soleimani et al. [70] proposed a "yes or no" question answering framework with two steps for finding particular individuals conducting one or several actions within aerial pictures. For the detection of multioriented vehicles within aerial images and videos, Li et al. [71] developed a rotatable region-based residual network (R^3-Net). To tackle the small-sized pedestrian problems, Xie et al. [72] proposed a context-aware pedestrian detection approach, i.e., deconvolution integrated faster R-CNN (DIF R-CNN), to integrate the deconvolutional module into DIF R-CNN for acquiring additional context information. Yang et al. [73] developed a clustered detection (ClusDet) network to unify the detection and clustering of the object within an end-to-end framework, covering a dedicated detection network (DetecNet), a scale estimation subnetwork (ScaleNet), as well as a cluster proposal subnetwork (CPNet). To solve the 1) large object size variation and 2) nonuniform object distribution problems, Li et al. [74] proposed a density-map-guided object detection network (DMNet), which involves a density map generation module, an image cropping module and an object detector. Liu et al. [75] proposed a high-resolution detection network (HRDNet) to take multiple resolutions with multidepth backbones as inputs. HRDNet involves a multiscale feature pyramid network (MS-FPN) and multidepth image pyramid network (MD-IPN) to optimize the detection of small objects and keep the performance of large-scale and middle-scale objects. Wu et al. [76] developed a dubbed nuisance disentangled feature transform (NDFT), which utilizes free meta-data with relevant UAV images to learn domain-robust features via an adversarial training framework.

B. Object Tracking

Object tracking can fall into two types: 1) generative tracking and 2) discriminative tracking. Generative tracking methods, such as Meanshift, Camshift, optical flow method, and particle filter, are capable of building a target model to extract target features and perform similar feature searches within subsequent frames. The discriminative model reveals that the target model and background information are both considered in the training process [77], [78]. The discriminative model acquires the target location in the current frame by comparing the differences between the background information and the target model. The discriminative model primarily has two directions: one is DCF-based methods, including MOSSE [79], CSK [80], KCF [81], and SAMF [82]; another is DL-based methods, such as MDNet [83], TCNN [84], and Siamese network [85]. Next, we review the representative discriminative tracker models for UAV object tracking [86], [87].

DCF-Based Tracker: Huang et al. [88] developed an aberrance repressed correlation filter (ARCF) to repress the aberrances in UAV object detection. By restricting the alteration rate in response maps generated at the detection phase, the ARCF tracker is capable of suppressing aberrances and exhibiting robustness and accuracy in tracking objects. Ye et al. [89] developed a multiregularized correlation filter (MRCF) through the regularization of the reliability of channels and the deviation of responses. The MRCF tracker can lead to adaptive channel weight distributions and smooth response changes simultaneously, which can effectively adapt to object appearance changes and enhance discriminability. In order to tackle the internal and external interference, Han proposed a state-aware anti-drift tracker (SAT) [90] by jointly learning the feature of the target and its surrounding patches. Afterward, Han et al. and Yuan et al. [91], [92] proposed several spatial-temporal context-aware tracking algorithms in accordance with DCF. Specifically, these models can learn a spatial-temporal context weight so that the target and background can be precisely distinguished under the UAV-tracking conditions. Furthermore, considering the aerial view and the small object scale under UAV-tracking scenarios, both of these DCF-based trackers incorporate the spatial context information to reduce background interference. Li et al. [93] proposed a spatially local response map change as spatial regularization, capable of learning spatio-temporal regularization terms online adaptively and automatically. Targeting the UAV tracking at night, Ye et al. [42] proposed a spatial-channel transformer-based low-light enhancer (SCT), which is trained based on the inspiration of a new task. Specifically, they developed a novel spatial-channel attention module for modeling information worldwide and retaining local context. During the enhancement process, SCT simultaneously denoises and illuminates nighttime images based on a nonlinear curve projection. Fu et al. [94] proposed a novel tracker learned by dynamic regression with automatic distractor repression (DR-Track), where the regression label is controlled dynamically for repressing distractors indicated as the local maximums. Yang et al. [95] transformed the large-scale least-squares problem in the spatial domain into several small-scale problems with constraints in the Fourier domain, using the correlation filter method to solve the real-time problems in UAV tracking.

DL-Based Tracker: The emergence of deep learning has brought a significant leap forward in visual object tracking filed, especially for the out-door scenes. Zhang et al. [96] proposed a coarse-to-fine deep scheme for tackling the ratio change problem in UAV tracking. First, the coarse-tracker generates an initial estimation of the target object, and then a sequence of actions is learned for fine-tuning the four boundaries of the bounding box. Jiang et al. [39] proposed a dual-flow semantic consistency (DFSC) method for UAV tracking. Under the modulation by the semantic flow across video sequences, the tracker can learn more robust class-level semantic information and obtain more discriminative instance-level features. To tackle the multiobjects tracking problem in UAV videos, Yu et al. [97] proposed a Siamese network to estimate global motion information in UAV video, which leverages the conditional generative adversarial networks (GAN) to produce the final motion prediction. Han et al. [12] combined the efficient DCF-based tracker with the precise DL model to eliminate the accumulating drift for the vehicle tracking. To be specific, the prediction for DCF tracker is incorporated as the input for a boundary regressing network, which are designed to correct the target's boundary, aiming at achieving a long-term tracking. Siamese models are employed to verify hand signature first [98], [99], and are gradually extended to object tracking task. Thanks to the powerful feature representation capability for CNNs, Siamese models present a great potential, as concluded in relevant surveys [100], [101], [102], [103], [104]. Although DL-based trackers could accomplish higher performance, it still face the difficulties in deploying efficient GPUs due to the limited size and computation resource on UAV platform. After all, the interference time for DL-based algorithms is relatively long, which could not meet the real-time standard for aerial tracking. Fu et al. [23] reviewed the research progress of Siamese trackers [105] and the development for high-performance embedded devices [106], pointed out a potential direction for Siamese UAV tracking.

C. Semantic Segmentation

Semantic segmentation aims to associate a label or category with each pixel in an image and identify collections of pixels that constitute different types [107], [108], [109]. There are two main types of semantic segmentation research: 1) the probabilistic graph model, such as [45] and 2) the DL-based methods [110] that have emerged over the past few years.

The probabilistic graph model, such as Markov random fields (MRF) and conditional random fields (CRF), establishes a probabilistic model with a graph to express the conditional dependence structure between random variables. It can model the joint probability distribution of the related image entities to perform semantic segmentation. At the same time, the rapid development of deep learning in computer vision also provides a basis for its application in remote sensing imagery [111]. The progress of the convolutional neural network in the pixel-by-pixel classification of images is based on massive data, such as Pascal VOC [112] and MS-COCO, in daily scenes. The remote sensing images are different from the daily scene ones with the characteristics of high spatial resolution, complex scenes, and numerous targets. Since then, more and more research has focused on applying CNNs to various remote sensing tasks. In the following, we mainly review the semantic segmentation of UAV images [86], [87] with a probabilistic graph model and DL-based methods.

Probabilistic Model: Yao et al. [113] constructed a triple-multipyramid structure, which combines the multiresolution, multiregion adjacency graph (RAG), and multisemantic elements. Kong et al. [114] exploited the geographical information of the region of interest in the form of a digital surface model (DSM) for urban UAV images semantic segmentation, which combines the visual features, DSM information, and a multiscale strategy with attention to improve the segmenting results.

DL-based Model: Sherrah et al. [110] proposed a deep fully convolution networks (FCN) without downsampling to obviate the need for deconvolution or interpolation. To more effectively exploit image features, they fine-tune the pretrained CNN on remote sensing data with a hybrid network. Kampffmeyer et al. [115] targeted the class imbalance problem for small objects. They use recent uncertainty measurement advances in CNNs and assess their qualitative and quantitative quality in a remote sensing context. Specifically, they adopt different deep architectures to cover the patch-based and so-called pixel-to-pixel methods and their integration for semantic segmentation. Maggiori et al. [116] derived a CNN framework adapted to the semantic segmentation problem, which can learn features at different resolutions and learn how to combine the above features. Girisha et al. [117] created a novel semantic segmentation dataset annotated manually. Moreover, they explore the performance of semantic segmentation algorithms for aerial videos achieved with the FCN and U-net architectures. Girisha et al. [118] proposed an enhanced encoder–decoder-based CNN architecture (UVid-Net) for UAV video semantic segmentation. The encoder can embed temporal information in terms of temporally consistent labeling. The decoder introduces the feature-refiner module to improve the location of the class labels.

Besides the semantic segmentation of images, video segmentation aims to divide pixels with consistent appearance and motion in video frames into continuous spatio-temporal communities. Video segmentation can be brought into remote sensing applications as a preprocessing module for further high-level applications. However, research on remote sensing video segmentation is extremely rare. Cheng et al. [119] developed a video segmentation algorithm by an expert mixture for aerial surveillance video. They employ trainable sequence maximum posterior probability for supervised image segmentation algorithm, mean-shift unsupervised image segmentation algorithm, and moving object detection algorithm. With the domain knowledge of aerial video surveillance, the outputs of the above three experts can be effectively combined to generate the final segmentation result. Teutsch assessed various object segmentation methods according to machine learning [120], blob extraction, and contour extraction. They proposed a local sliding window method with an AdaBoost classifier and integrated channel features. Wang proposed the S-MRF approach [45], which is a principled combination of superpixel labeling priors and the Markov random field for UAV semantic segmentation. Specifically, S-MRF utilizes the UAV metadata for motion estimation, followed by the superpixel labeling prior and MRF optimization.

IV. Evaluation Metrics

In addition, we need the evaluation metrics to quantitatively demonstrate the effectiveness of the object detection, object tracking, and semantic segmentation methods. In the following, we introduce some of the most commonly used evaluation metrics in these tasks.

A. Object Detection

We can measure the object detection methods from three aspects: 1) localization accuracy, 2) classification accuracy, and 3) efficiency.

Localization Accuracy: IoU is the most commonly used metrics, which calculates the ratio of the intersection and union of two sets of true and predicted values, generally represented as

$$\text{IoU} = \frac{TP}{FP + FN + TP}, \tag{1}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

Classification Accuracy: There are a lot of metrics such as Accuracy, Confusion Matrix, Precision, Recall, and AP.

Accuracy is defined as the correct predicted samples divided by the total samples

$$\text{Accuracy} = \frac{TP + TN}{FP + FN + TP + TN}. \tag{2}$$

Precision is defined as the ratio of the true positive samples in the data predicted as positive samples

$$\text{Precision} = \frac{TP}{FP + TP}. \tag{3}$$

Recall always accompanies Accuracy, which calculate the ratio between the predicted positive samples and total positive samples

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4}$$

Usually, the Precision–Recall curve is used in the object detection task to show the tradeoff between precision and recall of the classification.

Average precision (AP) and mean average precision (mAP) are another two important metrics in object detection algorithms. AP is the area under the Precision–Recall curve. mAP is the average of multiple class APs. For both AP and mAP, the higher, the better.

F1-score is the harmonic mean of precision and recall, which is calculated as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{5}$$

Receiver operating characteristic (ROC) is another common used metric, in which *x*-axis and *y*-axis represent FPR and TPR, respectively. When the TPR is larger but the FPR is smaller, the classification result is better.

Efficiency: FPS is always used to measure how many images are processed per second. The larger, the faster. Also, some works also measure the memory usage during the running time.

B. Object Tracking

Generally, researchers employ one-pass-evaluation (OPE) methodology [121], [122] to validate the accuracy and robustness for SOT algorithms. Each comparison trackers are initialized with the target's state (location and scale) given at the first frame of the video. Afterward, the tracking result is recorded for each subsequent frame no matter the tracker is located on the target or not. Based upon the OPE manner, two metrics (precision and success rate) are incorporated to evaluate the performance of comparison methods.

Precision Rate: Precision rate illustrates the percentage of the frames whose center location error (CLE) are within the given threshold between the predicted center for the candidate tracker C^{pr} with the one for the annotated bounding-box C^{bb}. Generally, 20 pixel is set as the threshold to determine whether the tracker is drift in each frame. However, for some specific scenarios, the threshold may be adjusted according to the target size. Since the precision rate varies across different videos, researchers generally average the precision score for all the sequences to obtain a comprehensive evaluation for the participated tracking algorithm on a certain dataset. It should be also noted that the precision metric can be easily affected by the image resolution and the bounding box scale, the normalized precision metric is also employed for performance evaluation in some literature by normalized the center location error over the scale for the bounding box

$$\text{CLE} = ||C^{bb} - C^{pr}||. \tag{6}$$

Success Rate: Success Rate is based upon the overlap ratio OP, which is defined as the intersection over union for the area of the annotated bounding box A^{bb} and the predicted one for candidate tracker A^{pr}. The success plot shows the percentage of the frames, where the overlap ratio is larger than a predefined threshold. In this way, we could obtain a continuous curve by linking the success rate under different threshold and the area under the curve (AUC) could be served as the second measure metric to rank the trackers

$$OP = \frac{A^{pr} \cap A^{bb}}{A^{pr} \cup A^{bb}}. \tag{7}$$

C. Sematic Segmentation

Pixel Accuracy (PA): PA represents the ratio of correct predictions for all pixel classes to the total number of pixels

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{8}$$

where $(k+1)$ is the total categories with k foreground categories and 1 background category. p_{ij} represents that the pixel of class i is predicted to be class j. When $i=j$, the prediction is correct, otherwise the prediction is wrong.

Mean Pixel Accuracy (MPA): Different from PA, MPA calculates the ratio of correct predictions to the total number of pixels in that category, then average the results for all categories

$$\text{MPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}. \tag{9}$$

Mean Intersection over Union (MIoU): In semantic segmentation, MIoU can be represented as the mean of the IoU among all categories

$$\text{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{10}$$

Frequency Weighted Intersection over Union (FWIoU): FWIoU is an improved version of MIoU. The difference between FWIoU and MIoU is the weighting way. MIoU applies the same weight $\frac{1}{k+1}$ to each category, while FWIoU uses the ratio between the number of each category and the total number as different weights for different categories

$$\text{MIoU} = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}\sum_{i=0}^{k}\frac{\sum_{j=0}^{k} p_{ij}p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{11}$$

V. Upcoming Domains and Future Opportunities

This work reviews several popular UAV task-related datasets and methods for Earth observation. This section summarizes some potential future directions for the UAV vision area.

First, the recent work lacks systematic validation. Most UAV vision tasks only rely on a certain or few datasets to validate the performance of their methods. They did not evaluate their methods on extensive datasets and for the various characteristics of the UAVs. Therefore, establishing a benchmark for evaluating different methods on extensive datasets for various characteristics of UAVs is a very useful direction in the future.

Second, real time is a significant problem in the UAV vision area. In recent years, deep learning has become a popular method for dealing with visual tasks because of its powerful recognition ability. However, it always requires a lot of computing resources. On the other hand, small UAVs cannot load with large devices such as GPUs. Therefore, how to perform real-time visual tasks on small devices is an urgent problem to be solved. Current researches focus on how to do vision tasks on the data captured by UAVs, and few works consider real-time problems.

Third, the technological advances in the UAV visual field are extending our capability at a breakneck speed, enabling many other data modalities of individual image data to be taken. For instance, recent development in imaging provides the opportunity to analyze infrared or other data modalities from different sensor devices, holding great promises to further transform the UAV visual field. Given the different modalities of such data (e.g., visible light, infrared), we can merge them before applying them to our tasks and anticipate UAV visual techniques to be readily adopted for the above data types when they become more available.

VI. Conclusion

The explosion of UAVs over the past few years has resulted in a resurgence in designing and employing the corresponding vision techniques for analyzing UAV data. In this study, we revisited and summarized the datasets for object detection, object tracking, and semantic segmentation methods for UAVs in the last decade. Subsequently, we reviewed the recent literature for their applications, summarized the achievements, and identified the missing aspects. Finally, we provide several research directions and practical considerations that we hope will spark future research in the application of the UAV vision era, such as the comprehensive study, real-time problem, and multimodality information.

References

[1] N. Audebert, B. Le Saux, and S. Lefevre, "How useful is region-based classification of remote sensing images in a deep learning framework?" in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 5091–5094.

[2] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[3] D. Honget al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[4] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.

[5] D. Honget al., "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021, Art. no. 5518615.

[6] W. Wang, Y. Han, C. Deng, and Z. Li, "Hyperspectral image classification via deep structure dictionary learning," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2266.

[7] H. Shi, Z. Fang, Y. Wang, and L. Chen, "An adaptive sample assignment strategy based on feature enhancement for ship detection in SAR images," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2238.

[8] H. Shi, Q. Sheng, Y. Wang, B. Yue, and L. Chen, "Dynamic range compression self-adaption method for SAR image based on deep learning," *Remote Sens.*, vol. 14, no. 10, 2022, Art. no. 2338.

[9] C. Deng, D. Jing, Y. Han, S. Wang, and H. Wang, "FAR-Net: Fast anchor refining for arbitrary-oriented object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art. no. 6505805.

[10] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, "A scale-aware pyramid network for multi-scale object detection in SAR images," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 973.

[11] B. Zhao, Y. Han, H. Wang, L. Tang, and T. Wang, "Robust shadow tracking for video SAR," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 821–825, May 2021.

[12] Y. Han, H. Wang, Z. Zhang, and W. Wang, "Boundary-aware vehicle tracking upon UAV," *Electron. Lett.*, vol. 56, no. 17, pp. 873–876, 2020.

[13] B. Zhao, H. Wang, L. Tang, and Y. Han, "Towards long-term UAV object tracking via effective feature matching," *Electron. Lett.*, 2020.

[14] C. Deng, S. He, Y. Han, and B. Zhao, "Learning dynamic spatial-temporal regularization for UAV object tracking," *IEEE Signal Process. Lett.*, vol. 28, pp. 1230–1234, 2021.

[15] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.

[16] L. Jiaoet al., "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[17] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[18] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3388–3415, Oct. 2021.

[19] W. Liet al., "Deep domain adaptive object detection: A survey," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2020, pp. 1808–1813.

[20] J. Chen, Q. Wu, D. Liu, and T. Xu, "Foreground-background imbalance problem in deep object detectors: A review," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2020, pp. 285–290.

[21] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: A survey," *Image Vis. Comput.*, vol. 104, 2020, Art. no. 104046.

[22] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 91–124, Mar. 2022.

[23] C. Fu, K. Lu, G. Zheng, J. Ye, Z. Cao, and B. Li, "Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis," 2022, *arXiv:2205.04281v2*.

[24] M. Barekatainet al., "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 28–35.

[25] P. Zhuet al., "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2021.

[26] M. Mandal, L. K. Kumar, and S. K. Vipparthi, "MOR-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2626–2635.

[27] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4145–4153.

[28] I. Bozcan and E. Kayacan, "Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 8504–8510.

[29] W. Zhang, C. Liu, F. Chang, and Y. Song, "Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1760.

[30] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, 2021.

[31] E. Bondiet al., "BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1736–1745.

[32] H. Zhang, M. Sun, Q. Li, L. Liu, M. Liu, and Y. Ji, "An empirical study of multi-scale object detection in high resolution UAV images," *Neurocomputing*, vol. 421, pp. 173–182, 2021.

[33] J. Wang, X. Teng, Z. Li, Q. Yu, Y. Bian, and J. Wei, "VSAI: A multi-view dataset for vehicle detection in complex scenarios using aerial images," *Drones*, vol. 6, no. 7, 2022, Art. no. 161.

[34] D. Duet al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.

[35] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[36] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 2118–2125.

[37] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4140–4146.

[38] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461.

[39] N. Jianget al., "Anti-UAV: A large multi-modal benchmark for UAV tracking," 2021, *arXiv:2101.08466*.

[40] C. Liuet al., "Aggregation signature for small object tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 1738–1747, 2020.

[41] B. Li, C. Fu, F. Ding, J. Ye, and F. Lin, "All-day object tracking for unmanned aerial vehicle," *IEEE Trans. Mobile Comput.*, to be published, doi: 10.1109/TMC.2022.3162892.

[42] J. Ye, C. Fu, Z. Cao, S. An, G. Zheng, and B. Li, "Tracker meets night: A transformer enhancer for UAV tracking," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3866–3873, Apr. 2022.

[43] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, "Siamese anchor proposal network for high-speed aerial tracking," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 1–7.

[44] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, "Onboard real-time aerial tracking with efficient Siamese anchor proposal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art no. 5606913.

[45] Y. Wang, W. Ding, B. Zhang, H. Li, and S. Liu, "Superpixel labeling priors and MRF for aerial video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2590–2603, Aug. 2020.

[46] Y. Lyu, G. Vosselman, G. Xia, A. Yilmaz, and M. Y. Yang, "The UAVid dataset for video semantic segmentation," 2018, *arXiv:1810.10438*.

[47] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1499–1508.

[48] S. Girisha, M. M. Pai, U. Verma, and R. M. Pai, "Performance analysis of semantic segmentation algorithms for finely annotated new UAV aerial video dataset (manipaluavid)," *IEEE Access*, vol. 7, pp. 136239–136253, 2019.

[49] G.-S. Xiaet al., "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[50] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.

[51] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016.

[52] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory prediction in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 549–565.

[53] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[55] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[57] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[58] W. Liuet al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[59] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[60] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[61] C. Chenet al., "RRNet: A hybrid detector for object detection in drone-captured images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 100–108.

[62] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in UAV vision based on cascade network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 118–126.

[63] Y. Hu, X. Wu, G. Zheng, and X. Liu, "Object detection of UAV for anti-UAV based on improved YOLO v3," in *Proc. Chin. Control Conf.*, 2019, pp. 8386–8390.

[64] J. Deng, Z. Shi, and C. Zhuo, "Energy-efficient real-time UAV object detection on embedded platforms," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 39, no. 10, pp. 3123–3127, Oct. 2020.

[65] P. Zhang, Y. Zhong, and X. Li, "SlimYOLOv3: Narrower, faster and better for real-time UAV applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 37–45.

[66] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, Jun. 2019.

[67] Y. Liu, Z. Ding, Y. Cao, and M. Chang, "Multi-scale feature fusion UAV image object detection method based on dilated convolution and attention mechanism," in *Proc. 8th Int. Conf. Inf. Technol., IoT Smart City*, 2020, pp. 125–132.

[68] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2238.

[69] Z. Li, X. Liu, Y. Zhao, B. Liu, Z. Huang, and R. Hong, "A lightweight multi-scale aggregated model for detecting aerial images captured by UAVs," *J. Vis. Commun. Image Representation*, vol. 77, 2021, Art. no. 103058.

[70] A. Soleimani and N. M. Nasrabadi, "Convolutional neural networks for aerial multi-label pedestrian detection," in *Proc. 21st Int. Conf. Inf. Fusion*, 2018, pp. 1005–1010.

[71] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R^3-Net: A deep network for multi-oriented vehicle detection in aerial images and videos," 2018, *arXiv:1808.05560*.

[72] H. Xie, Y. Chen, and H. Shin, "Context-aware pedestrian detection especially for small-sized instances with deconvolution integrated faster RCNN (DIF R-CNN)," *Appl. Intell.*, vol. 49, no. 3, pp. 1200–1211, 2019.

[73] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8311–8320.

[74] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 190–191.

[75] Z. Liu, G. Gao, L. Sun, and Z. Fang, "Hrdnet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.

[76] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon, and Z. Wang, "Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1201–1210.

[77] C. Deng, Y. Han, and B. Zhao, "High-performance visual tracking with extreme learning machine framework," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2781–2792, Jun. 2020.

[78] Z. Zhao, Y. Han, T. Xu, X. Li, H. Song, and J. Luo, "A reliable and real-time tracking method with color distribution," *Sensors*, vol. 17, no. 10, 2017, Art. no. 2303.

[79] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2544–2550.

[80] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.

[81] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2014.

[82] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 254–265.

[83] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6428–6436.

[84] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," 2016, *arXiv:1608.07242*.

[85] Z. Zhang and H. Peng, "Deeper and wider Siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4591–4600.

[86] S. Zhang, L. Zhuo, H. Zhang, and J. Li, "Object tracking in unmanned aerial vehicle videos via multifeature discrimination and instance-aware attention network," *Remote Sens.*, vol. 12, no. 16, 2020, Art. no. 2646.

[87] S. Kapania, D. Saini, S. Goyal, N. Thakur, R. Jain, and P. Nagrath, "Multi object tracking with UAVs using deep sort and YOLOv3 retinanet detection framework," in *Proc. 1st ACM Workshop Auton. Intell. Mobile Syst.*, 2020, pp. 1–6.

[88] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time UAV tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2891–2900.

[89] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, "Multi-regularized correlation filter for UAV tracking and self-localization," *IEEE Trans. Ind. Electron.*, vol. 69, no. 6, pp. 6004–6014, Jun. 2021.

[90] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.

[91] Y. Han, C. Deng, B. Zhao, and B. Zhao, "Spatial-temporal context-aware tracking," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 500–504, Mar. 2019.

[92] D. Yuan, X. Chang, Z. Li, and Z. He, "Learning adaptive spatial-temporal context-aware correlation filters for UAV tracking," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3, pp. 1–18, 2022.

[93] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11923–11932.

[94] C. Fu, F. Ding, Y. Li, J. Jin, and C. Feng, "Learning dynamic regression with automatic distractor repression for real-time UAV tracking," *Eng. Appl. Artif. Intell.*, vol. 98, 2021, Art. no. 104116.

[95] Y. Xiaoyuan, Z. Ridong, W. Jingkai, and L. Zhengze, "Real-time object tracking via least squares transformation in spatial and fourier domains for unmanned aerial vehicles," *Chin. J. Aeronaut.*, vol. 32, no. 7, pp. 1716–1726, 2019.

[96] W. Zhang, K. Song, X. Rong, and Y. Li, "Coarse-to-fine UAV target tracking with deep reinforcement learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1522–1530, Oct. 2018.

[97] H. Yu, G. Li, L. Su, B. Zhong, H. Yao, and Q. Huang, "Conditional GAN based individual and global motion fusion for multiple object tracking in UAV videos," *Pattern Recognit. Lett.*, vol. 131, pp. 219–226, 2020.

[98] SHAH ROOPAKet al., "Signature verification using a "Siamese" time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 07, no. 4, pp. 669–669, 1993.

[99] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1993, pp. 737–744.

[100] S. M. Marvasti-Zadeh, C. Li, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 3943–3968, May 2022.

[101] M. Ondrasovic and P. Tarabek, "Siamese visual object tracking: A survey," *IEEE Access*, vol. 9, pp. 110149–110172, 2021.

[102] R. Pflugfelder, "An in-depth analysis of visual tracking with Siamese neural networks," 2017, *arXiv:1707.00569*.

[103] S. Javed, M. Danelljan, F. S. Khan, M. H. Khan, M. Felsberg, and J. Matas, "Visual object tracking with discriminative filters and Siamese networks: A survey and outlook," to be published, doi: 10.1109/TPAMI.2022.3212594.

[104] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 33–40.

[105] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "SiamAPN++: Siamese attentional aggregation network for real-time UAV tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 3086–3092.

[106] S. Mittal, "A survey on optimized implementation of deep learning models on the Nvidia Jetson platform," *J. Syst. Architect.*, vol. 97, pp. 428–442, 2019.

[107] H. Shi, J. Fan, Y. Wang, and L. Chen, "Dual attention feature fusion and adaptive context for accurate segmentation of very high-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3715.

[108] Y. Wang, H. Shi, S. Dong, Y. Zhuang, and L. Chen, "Dual-path sparse hierarchical network for semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021, Art. no. 8010505.

[109] T. Wei, J. Wang, W. Liu, H. Chen, and H. Shi, "Marginal center loss for deep remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 968–972, Jun. 2019.

[110] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.

[111] C. Liuet al., "RB-Net: Training highly accurate and efficient binary neural networks with reshaped point-wise convolution and balanced activation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6414–6424, Sep. 2022.

[112] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[113] H. Yao, X. Wang, L. Zhao, M. Tian, L. Gong, and B. Li, "Semantic segmentation for remote sensing images using pyramid object-based markov random field with dual-track information transmission," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021, Art. no. 8023105.

[114] Y. Kong, B. Zhang, B. Yan, Y. Liu, H. Leung, and X. Peng, "Affiliated fusion conditional random field for urban UAV image semantic segmentation," *Sensors*, vol. 20, no. 4, 2020, Art. no. 993.

[115] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1–9.

[116] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “High-resolution aerial image labeling with convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

[117] S. Girisha, M. P. MM, U. Verma, and R. M. Pai, “Semantic segmentation of UAV aerial videos using convolutional neural networks,” in *Proc. IEEE Second Int. Conf. Artif. Intell. Knowl. Eng.*, 2019, pp. 21–27.

[118] S. Girisha, U. Verma, M. M. Pai, and R. M. Pai, “UVID-Net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4115–4127, 2021.

[119] H. Cheng and D. Butler, “Segmentation of aerial surveillance video using a mixture of experts,” in *Proc. Digit. Image Comput.: Techn. Appl.*, 2005, pp. 66–66.

[120] M. Teutsch, W. Krüger, and J. Beyerer, “Evaluation of object segmentation to improve moving vehicle detection in aerial videos,” in *Proc. IEEE 11th Int. Conf. Adv. Video Signal Based Surveill.*, 2014, pp. 265–270.

[121] Y. Wu, J. Lim, and M.-H. Yang, “Online object tracking: A benchmark,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2411–2418.

[122] Y. Wu, J. Lim, and M. H. Yang, “Object tracking benchmark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

Yuqi Han received the B.Eng. degree in information engineering from Beijing Institute of Technology, Beijing, China, the B.Sc. degree in the field of Economy from National School of Development, Peking University, Beijing, China, in 2015, and the Ph.D. degree in information and communication engineering with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2021.

He is currently a Research Fellow with the Department of Computer Science and Technology, Tsinghua Univeristy, Beijing, China. His research interests include computer vision, remote sensing and UAV.

Huaping Liu (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2004.

He is currently a Tenured Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He is also with the State Key Laboratory of Intelligent Systems and Technology and the Beijing National Research Center for Information Science and Technology, Beijing, China. In 2020, he was recognized as a Distinguished Young Scholar by the Natural Science Foundation of China. His research interests include robotics, dynamic systems, and machine learning, with particular emphasis on robotic perception, learning, and control.

Yufeng Wang (Member, IEEE) received the B.S. degree in communication engineering from Northwestern Polytechnical University, Xi’an, China, in 2015, and the Ph.D. degree in information and communication system from Beihang University, Beijing, China, in 2021.

He has been with the Institute of Unmanned System, Beihang University, since 2021. His research interests include computer vision and machine learning.

Chunlei Liu received the B.S. degree in information engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016, and the Ph.D. degree in information and communication system with the Department of Electrical and Information Engineering, Beihang University, Beijing, China, in 2022.

She is currently a Postdoc with the Children Medical Research Institute, Univerisity of Sydney, Parramatta, NSW, Australia. Her research interests include computer vision, machine learning, and pattern recognition.