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.

The MONET dataset: Multimodal drone thermal dataset recorded in rural scenarios

Luigi Riz[1] Andrea Caraffa[1] Matteo Bortolon[1] Mohamed Lamine Mekhalfi[1] Davide Boscaini[1]
André Moura[2] José Antunes[2] André Dias[2] Hugo Silva[2] Andreas Leonidou[3]
Christos Constantinides[3] Christos Keleshis[3] Dante Abate[3] Fabio Poiesi[1]

[1]Fondazione Bruno Kessler [2]INESC TEC [3]The Cyprus Institute

Abstract

We present MONET, a new multimodal dataset captured using a thermal camera mounted on a drone that flew over rural areas, and recorded human and vehicle activities. We captured MONET to study the problem of object localisation and behaviour understanding of targets undergoing large-scale variations and being recorded from different and moving viewpoints. Target activities occur in two different land sites, each with unique scene structures and cluttered backgrounds. MONET consists of approximately 53K images featuring 162K manually annotated bounding boxes. Each image is timestamp-aligned with drone metadata that includes information about attitudes, speed, altitude, and GPS coordinates. MONET is different from previous thermal drone datasets because it features multimodal data, including rural scenes captured with thermal cameras containing both person and vehicle targets, along with trajectory information and metadata. We assessed the difficulty of the dataset in terms of transfer learning between the two sites and evaluated nine object detection algorithms to identify the open challenges associated with this type of data. Project page: https://github.com/fabiopoiesi/monet_dataset.

1. Introduction

Thermal image understanding enables localisation of objects that may not be visible through traditional RGB cameras. This can be useful in a variety of applications, such as in surveillance and security, where illicit activities typically occur overnight [16, 25], or in search and rescue [4], and military operations, where targets can be easier to locate based on their emitted heat rather than their cloth textures.

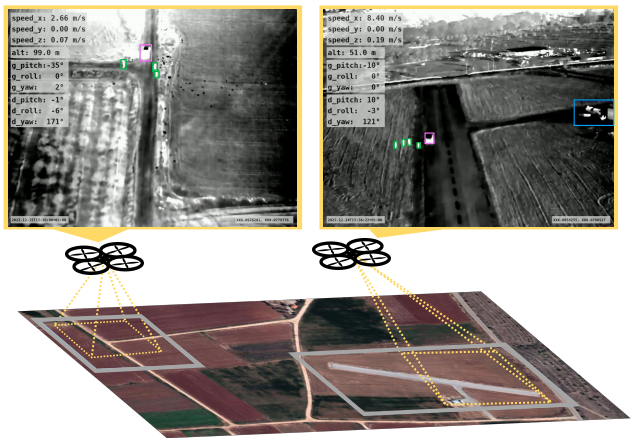


Figure 1. The MONET dataset is captured with a thermal camera mounted on a multirotor drone. Three types of objects are annotated with bounding boxes: green is `person`, magenta is `vehicle`, and blue is `ignore`. Outlined in grey are the two recording sites: left-hand site is *dirt-road* and the right-hand site is *runway*. MONET provides drone metadata (e.g. attitude, GPS, altitude) timestamp aligned to each image. The map below is a screenshot from Google Maps. Outlined in yellow are the two fields of view associated to the example thermal images.

Object localisation in thermal images presents several challenges. Typically, thermal imaging cameras have lower resolution than traditional RGB cameras, hence distinguishing fine-scale details in images is difficult. Thermal imaging cameras are susceptible to noise and ghost effects, such as interference from other sources of heat or surface reflections. The appearance of objects can vary depending on their temperature and emissivity, as well as the ambient temperature and humidity, hence algorithms should be robust to different environment conditions. Thermal images can be cluttered, thus foreground objects may be indistinguishable from their background. Similarly to RGB images, objects

This work was carried out within the scope of the SHIELD project that received funding from the European Union's Joint Programming Initiative – Cultural Heritage, Conservation, Protection and Use joint call.

can also occlude each other, thus hindering multiple object localisation. Objects of the same size and with the same temperature (e.g. a boulder and a person), if captured from a distance, may appear with a similar silhouette. In order to develop algorithms for object detection in thermal images, it is necessary to have a large and diverse dataset of annotated images. However, such datasets may be difficult to obtain, because collecting large amounts of thermal data from drones and annotating them is often costly.

Publicly, we can find several drone datasets recorded with visible spectrum cameras [3, 8, 9, 20, 28], instead thermal datasets are less popular. There are some thermal datasets that are either annotated for single-object tracking applications [1, 14] (one target max per frame is annotated), or they are captured from static cameras resembling images captured from aerial vehicles [18, 25]. Our focus is understanding scenes that can potentially contain multiple objects and that are recorded from moving drones, but only few datasets are available with such desired properties [2, 22]. One is BIRDSAI [2], a long-wave thermal infrared (LWIR) dataset that contains nighttime images of animals and humans in Southern Africa. Another one is HIT-UAV [22], a LWIR dataset that contains both nighttime and daytime images of humans, bicycles, and vehicles, captured by a drone in urban scenarios (schools, parking lots, roads, playgrounds) flying between 60m to 130m altitude. The SeaDroneSee dataset [24] also provides thermal images captured from drones, but it includes scenes with humans and vehicles in water. Although water is a challenging scenario, its challenges are different from terrain scenarios, i.e. the structure of the environment is different, and background materials can make targets indistinguishable if they emit the same heat. Moreover, drone metadata (e.g. speed, altitude, drone and gimbal attitudes) is an important piece of information because one can use it to retrieve the expected scale of the targets to detect [17], or to calibrate the motion models of tracking algorithms [12]. To this end, SeaDroneSee provides a comprehensive list of metadata. HIT-UAV only provides information about altitude, camera perspective, and a day/night flag. BIRDSAI does not provide metadata.

In this paper, we present MONET, a multimodal drone thermal dataset recorded in rural scenarios that provides timestamp-aligned images and drone metadata (see Fig. 1). MONET comprises approximately 53K frames, with about 162K manually-annotated bounding boxes. The dataset includes two main target categories, i.e. `person` and `vehicle`, plus a third category for a region to `ignore`. Frames with targets contain 2.96 people and 1.33 vehicles on average. Target bounding boxes are annotated with identities for multi-object tracking applications. MONET's metadata includes drone and gimbal attitude (pitch, roll, and yaw), GPS, altitude, and speed (in the x, y, and z axes). The dataset presents challenges such as sudden and fast camera motion, background heat, large-scale variations, and different environmental structures. We define two dataset scenarios, i.e. *runway* and *dirt-road*, which include scenes recorded near a runway and in an agricultural land, respectively. Using the Faster R-CNN detector [19], we analyse MONET's challenges when training and evaluating on the same scenario (e.g. runway to runway) and on different scenarios (runway to dirt-road). Although the sensor is the same, experiments show that training on one scenario and evaluating on the other leads to a significant drop in performance. We evaluate nine object detectors and discuss and analyse MONET's challenges through qualitative results.

2. Related real-world datasets

We conduct a survey of related real-world datasets by analysing various factors, including (i) object categories such as people, vehicles, and animals, (ii) different sensor modalities, such as visible and infrared, (iii) different scenarios, such as urban, rural, and maritime, (iv) different camera views, such as fixed top-down and varying in different directions, and (v) different drone attitudes, such as static and moving. The datasets were recorded at various times of the day, with different weather conditions and at different altitudes, and all of them provide annotations in the form of bounding boxes (see Tab. 1).

The Campus dataset [20] comprises approximately 930K images with about 11M bounding box instances of pedestrians, bicyclists, and vehicles. These targets interact with each other within the Stanford University campus. The Campus dataset was designed to facilitate multi-object tracking, activity understanding, and trajectory forecasting. Object trajectories, along with their IDs, were annotated. The images were captured in the visible spectrum using a top-down camera during daytime from a multirotor drone hovering at an altitude of about 80m.

The Car Parking Lot Dataset (CARPK) [9] comprises approximately 1.5K images with about 90K bounding box instances of cars from four different parking lots in urban scenarios. CARPK was designed for car counting, and no target IDs were annotated. The images were captured in the visible spectrum using a top-down camera during daytime from a multirotor drone flying at an altitude of 40m.

The VisDrone dataset [28] comprises approximately 40K images with about 2.5M bounding box instances of pedestrians, vehicles, and bicycles, captured from 14 cities in China, between urban and rural scenarios. The images were captured in the visible spectrum with arbitrary camera viewpoints during both daytime and nighttime from a multirotor drone flying at different altitudes. VisDrone provides object trajectory annotations, but no altitude information.

The UAVDT dataset [8] comprises 80K images with about 2.7K vehicles and 841K bounding box vehicle instances, such as cars, trucks, and buses, from different urban

Table 1. Relevant publicly available real-world datasets recorded from drones outdoors. Keys. V: Visible. NIR: Near Infrared ($841nm$). RE: Red Edge ($717nm$). LWIR: Long-Wave Infrared (7.5-13.5μm). The dataset name is clickable and links to the dataset webpage.

Dataset attribute		**Campus** [20]	**CARPK** [9]	**VisDrone** [28]	**UAVDT** [8]	**BIRDSAI** [2]	**AU-AIR** [3]	**SeaDronesSee** [24]	**HIT-UAV** [22]	**MONET**
# images		930K	1.5K	40K	80K	62K	33K	54K	2.9K	53K
# bounding boxes		11M	90K	2.5M	841K	154K	132K	400K	25K	162K
# object trajectories		✓		✓	✓	✓		✓		✓
Metadata					✓		✓	✓	✓	✓
Categories	People	✓		✓		✓	✓	✓	✓	✓
	Vehicles	✓	✓	✓	✓		✓	✓	✓	✓
	Animals					✓				
Sensor modality		V	V	V	V	LWIR	V	V+NIR+RE	LWIR	LWIR
Time/Weather	Day/Clear	✓	✓	✓	✓		✓	✓	✓	✓
	Day/Foggy				✓					
	Night/Clear			✓	✓	✓			✓	✓
Scenario	Urban	✓	✓	✓	✓		✓		✓	
	Rural			✓		✓				✓
	Maritime							✓		
Camera view	Fixed	✓	✓				✓			
	Direction	top-down	top-down	varying	varying	varying	varying	varying	varying	varying
Drone attitude	Static	✓								
	Moving		✓	✓	✓	✓	✓	✓	✓	✓
Altitude [m]		80	40	n.a.	10-70	60-120	5-30	5-260	60-130	20-130
Year		2016	2017	2018	2018	2020	2020	2022	2022	2023

scenarios. The images were captured in the visible spectrum from arbitrary viewpoints during both daytime and nighttime using a multirotor drone at different altitudes. UAVDT provides object trajectory annotations and sequence-level metadata information, including i) time/weather conditions (daytime, nighttime, and fog), ii) flying altitude (low: 10-30m, medium: 30-70m, and high: above 70m), and iii) camera views (front, side, and bird).

The BIRDSAI dataset [2] comprises 62K images with about 120K animal and 34K human bounding box instances from different national parks in Southern Africa. BIRDSAI was designed for protected area monitoring to curb illegal activities like poaching and animal trafficking. BIRDSAI provides object trajectory annotations. The images were captured in the LWIR spectrum from arbitrary viewpoints during nighttime by using a fixed-wing drone flying at altitudes between 60m to 120m.

The AU-AIR dataset [3] comprises 33K images with about 132K bounding box instances of people and vehicles in an urban scenario. This dataset was designed for object detection tasks, hence target IDs are unavailable. The images were captured in the visible spectrum with arbitrary camera viewpoints during daytime by using a multirotor drone flying at altitudes between 5m to 30m. Unlike UAVDT, AU-AIR provides image-level metadata, which includes drone speed, roll, pitch, yaw, altitude, latitude, and longitude. The camera is fixed on the drone and it points in different directions during the flight.

The SeaDroneSee dataset [24] comprises 54K images with about 400K bounding box instances of people and vehicles (boats) in different maritime scenarios. SeaDroneSee was designed for search and rescue applications, specifically for benchmarking multi-object tracking algorithms, hence object trajectories are provided. Images were captured in both the visible and infrared spectrum by using fixed-wing and multirotor drones flying at altitudes between 5m to 260m. Similarly to AU-AIR, SeaDroneSee provides metadata logged at 10Hz, which include drone speed, attitude, altitude, GPS, and gimbal pitch.

The HIT-UAV dataset [22] comprises 2.9K images with about 25K bounding box instances of people and vehicles in urban scenarios. This dataset was designed for object detection tasks, hence target IDs are unavailable. Images were captured in the LWIR spectrum with arbitrary camera viewpoints during both daytime and nighttime by using a multirotor drone flying at altitudes between 60m to 130m.

Unlike BIRDSAI, MONET includes metadata and the vehicle category instead of the animal category, and is recorded from a multirotor drone as opposed to a fixed-wing drone. Unlike HIT-UAV, MONET includes more metadata, and has several more annotations that can also be used for multiple object-tracking applications.

3. Hardware

3.1. Multirotor drone

We used a fully-customised multirotor drone, which was designed for automated surveillance and detection of ar-

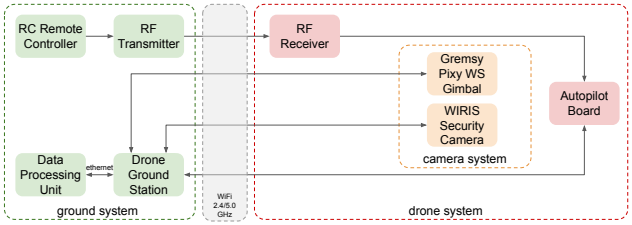


Figure 2. Data acquisition architecture that includes ground and drone system, and WiFi communication between these two.

chaeological looting activities. The drone includes the airframe, payload, propulsion, control, and communication systems. The airframe is designed to be compact, lightweight, and capable of carrying payloads up to 1.5kg. The propulsion system is composed of eight motors and eight Electronic Speed Controllers, ensuring stability and redundancy. It has four arms that support these eight motors. The control system includes a high-frequency IMU, an accurate barometric altimeter, and an external GPS and compass module. It supports several flight modes, including manual, stabilised, heading hold, hovering, automated waypoint navigation, return to home, auto take-off and auto-landing, and provides real-time monitoring through an on-screen display. The communication system provides real-time control and monitoring of the drone, and its payload.

3.2. Data acquisition system

The camera acquisition system consists of i) the camera WIRIS Security from Workswell[1] that features two separate sensors, i.e. RGB and thermal, ii) the gimbal unit to physical hold the camera on the drone, and iii) the Data Processing Unit. Fig. 2 illustrates the architecture.

The WIRIS Security is provided with a proprietary SDK. In order to control the camera, we developed a customised module that implements the WIRIS Security control commands (through SDK) to the camera via Ethernet connection, which is implemented in the drone ground station using ROS [21]. The RGB and thermal images are received in the Data Processing Unit as RTSP video streams, and published in ROS topics by using two independent nodes. The RGB sensor allows up to 30x optical zoom, Full HD (1920×1080) resolution, at a framerate of 30Hz. The thermal sensor operates in the LWIR spectrum (7.5-13.5 μm) featuring an 800×600 resolution with -20°C to 150°C thermal sensitivity. The camera includes an internal SSD drive with 256GB of storage space, allows communication through Ethernet, USB and HDMI.

[1] https://workswell-thermal-camera.com/drone-security-thermal-imaging-camera-night-vision-uav: last access: Apr. 2023.

4. Dataset

We collected MONET in a rural area near the city of Nicosia, Cyprus, in mid-December. The sequences were captured in the afternoon, evening and night. One recording site is on a runway, which is property of The Cyprus Institute, the other one is on agricultural lands. We name these sites as *runway* and *dirt-road*, respectively. Fig. 1 shows the two sites.

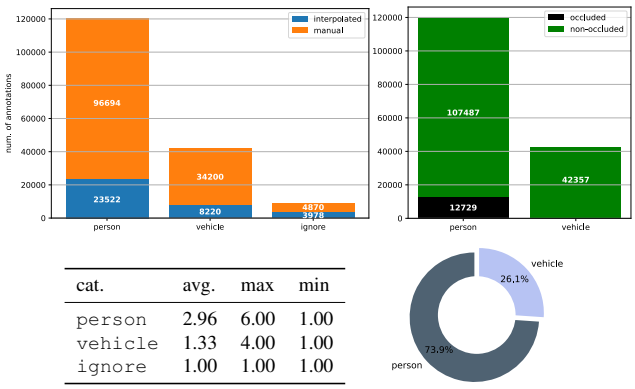
4.1. Annotation procedure

Six people contributed to the annotation of MONET by using CVAT [7]. CVAT was installed on a server and utilised via web browser. We created an account for each annotator. A certain number of distinct sequences were assigned to each annotator. We asked annotators to follow a set of guidelines: i) bounding boxes should be drawn as tight as possible on the targets as long as the object is clearly distinguishable from its background; ii) bounding box interpolation across frames is allowed as long as each frame is checked to see if the bounding boxes are correctly centred on targets; iii) brightness, contrast and saturation can be adjusted through CVAT UI to make targets more distinguishable; iv) if a target is partially occluded by another object or indistinguishable from the background, its bounding box should be drawn based on the best guess of the annotator and then flagged as *occluded*; v) the annotation of a target should start when more than ∼30% of its pixels are in the scene; vi) the annotation of a target should terminate when more than ∼70% of its pixels are outside the scene; vii) if a target exits the scene and then re-enters a new ID should be associated to it. Once the annotations were completed, three people double-checked them to ensure they were accurate and consistent with the guidelines. The Supplementary Material contains annotation examples.

4.2. Bounding box categories and statistics

We annotated three types of targets: `vehicle`, `person`, and `ignore`. The bounding boxes of `vehicle` include car-like objects, `person` is self-explanatory, and `ignore` include the hangar location next to the runway. We decided to ignore this location because other people and vehicles are often visible next to the hangar, and we want to avoid data-driven detection algorithms to learn pattern biases (e.g. people and vehicles often next to the hangar structure). Therefore, we zeroed the regions of the images defined by the `ignore` bounding boxes in order to avoid this bias during training. We provide the image data in its original form if one wants to exploit this location or additional targets for different purposes.

Fig. 3 illustrates the bounding box annotation distributions of the whole dataset. The left-hand side graph shows that the largest portion of bounding boxes is man-



cat.	avg.	max	min
person	2.96	6.00	1.00
vehicle	1.33	4.00	1.00
ignore	1.00	1.00	1.00

Figure 3. Bounding box annotation statistics of the MONET dataset. The top-left graph highlights the number of interpolated and manually drawn bounding boxes. The top-right graph highlights the number of occluded and non-occluded bounding boxes: vehicle features 63 occluded bounding boxes (ignore is omitted because always non-occluded). The table shows the average, maximum and minimum of number of targets that can be found. The doughnut chart summarises the percentage of person and vehicle bounding boxes.

ually drawn, while the others are (linearly) interpolated. Interpolation is a feature of CVAT and is applied between consecutive frames. While interpolation can be effective when annotations are made on videos captured from static cameras, this is not the case when the camera is moving. We only exploited interpolation occasionally because the motion of the drone plus the motion of the targets cannot be modelled with a linear motion model. The right-hand side graph shows that the largest portion of bounding boxes is visible, while the others are occluded. The annotators flag a bounding box as occluded when they deem the target was significantly occluded by another target, or when the target was indistinguishable from the background, e.g. due to interfering emitted temperatures. Although the target is indistinguishable in some frames, we purposely annotated the bounding boxes (flagging them as occluded) as it is a potential challenge if one's use case is tracking. We investigate the effect of training detectors with and without occluded bounding box in Sec. 5.

4.3. Drone metadata

Together with the camera frames, MONET also includes drone metadata information. As for the images, we logged the timestamp of each metadata during dataset acquisition. Metadata was captured at about 40Hz on average. Like [24], we used nearest-neighbour assignment between image and metadata timestamps. Metadata includes the date in ISO 8601 format, the drone and gimbal attitudes (pitch, roll, yaw), latitude, longitude, altitude and speed (x, y, and z axes). Tab. 2 includes a detailed list of the metadata we collected along with their minimum and maximum values.

Table 2. MONET metadata specifications indicating the units of each metadata along with its minimum and maximum value.

Data	Unit	min. value	max. value
data	ISO 8601	-	-
drone pitch	degrees	-90	90
drone roll	degrees	-90	90
drone yaw	degrees	0	360
gimbal pitch	degrees	-40	90
gimbal roll	degrees	-45	45
gimbal yaw	degrees	-180	180
latitude	degrees	-90	90
longitude	degrees	-180	180
altitude	m	0	user defined
x-axis speed	cm/s	0	2800
y-axis speed	cm/s	0	2800
z-axis speed	m/s	0	10

4.4. Examples of images and annotations

Fig. 4 shows examples of annotations in dirt-road and runway scenarios recorded from the altitudes of 80m (a-c) and 100m (d). In particular, Fig. 4a shows four person targets in the dirt-road scenario in normal visibility conditions. Fig. 4b shows the same four targets as before, but when one target is flagged as occluded: the small difference in the measured heat between target and background makes them nearly indistinguishable. Fig. 4c shows a similar case where a target is difficult to distinguish from the background, but in this case this is due the vehicle's heat behind the person. Lastly, we show an example of an ignore region. This is the hangar hosting the drone operators. We train our detection algorithms by zeroing the region of the image defined by the ignore bounding box. See the project page for videos of dirt-road and runway showing the annotations along with the aligned metadata.

5. Experiments

5.1. Experimental setup

Scenes and settings. We split the dataset into two scenes: *dirt-road* and *runway*, which include recordings of people activities nearby a runway and in an agriculture land, respectively. Dirt-road is composed of 23.3K frames with 83.4K bounding box annotations, while runway is composed of 29.4K frames with 79.3K bounding box annotations. Each scene is divided into disjoint splits for training, validation, and test. In Fig. 1 we can see that the structure of the environment is different between these two scenes. So we conduct two sets of experiments to analyse the challenges of MONET in terms of transfer learning when object detection algorithms are trained and evaluated on different

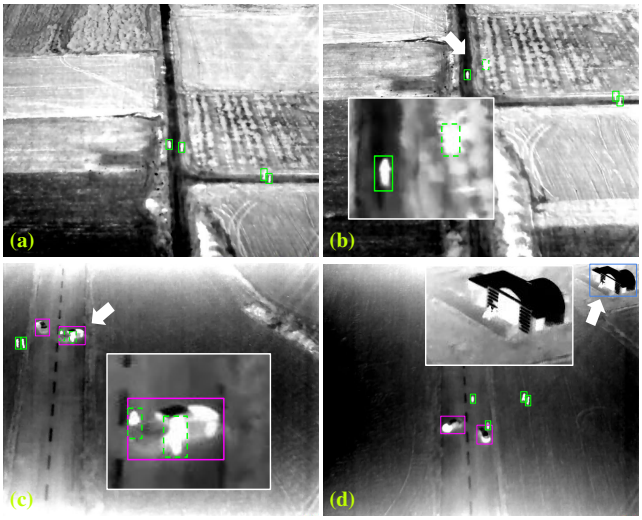


Figure 4. Annotation examples showing (a,b) dirt-road and (c,d) runway scenarios. Bounding boxes: green for `person`, magenta for `vehicle`, and blue for `ignore`. Recording altitudes (a-c) 80m and (d) 100m. Dotted bounding boxes indicate occlusions: (b) Example of occlusion where the target is indistinguishable from the background. (c) Example of occlusion where the heat from a vehicle makes the person indistinguishable. (d) Example of bounding box labelled as `ignore`, enclosing the hangar area.

scenes. Firstly, we focus on Faster R-CNN [19] and assess several transfer learning configurations, including the combination of MONET and HIT-UAV [22] data. Secondly, we compare the performance of nine popular object detectors, aiming to understand MONET's challenges. We use COCO evaluation procedure and report AP, AP_{50}, AP_{75}, and AP per class [13]. Because each detector is calibrated differently, setting a comparable detection confidence threshold is impractical. Therefore, we evaluate all the detectors by using confidence 10^{-8}. This aspect is further discussed in the Supplementary Material.

Detectors. We use the MMDetection open source object detection toolbox [6] for the following implementations: Faster R-CNN (2015) [19], SSD (2016) [15], CornerNet (2018) [11], FCOS (2019) [23], DETR (2020) [5], Deformable DETR (2021) [29], and VarifocalNet (2021) [27], while we use authors' implementation for ObjectBox (2022) [26] and YOLOv8 (2023) [10]. We train all these methods with the same data augmentations, and, where possible, with the same backbones and optimisation parameters. Only for ObjectBox and YOLOv8 we perform additional experiments with the data augmentations proposed by the authors to investigate if they lead to different results. Please refer to the Supplementary Material for the training configuration of each detector.

Table 3. Evaluation of Faster R-CNN initialised with COCO trained weights. Values are provided in percentage. Grey background indicates transfer learning results. Keys: V: validation split. T: test split. †: additional pre-training on top of COCO initialisation. w/o occ.: without bounding boxes with occlusion flag.

Exp	Train/Eval	Detector	AP	AP_{50}	AP_{75}	class-AP person	class-AP vehicle
1	dirt-road +runway	dirt-road V	24.9	63.8	6.5	10.9	39.0
		dirt-road T	36.3	82.4	22.0	33.1	39.6
		runway V	44.5	90.2	38.9	31.8	57.2
		runway T	42.1	84.0	37.3	37.8	46.5
2	dirt-road +runway (w/o occ.)	dirt-road V	25.1	67.8	7.8	13.6	36.6
		dirt-road T	36.3	81.3	21.9	32.1	40.5
		runway V	39.1	82.9	35.7	18.6	59.6
		runway T	47.6	88.5	46.0	43.2	51.9
3	HIT-UAV† +dirt-road +runway	dirt-road V	28.4	70.1	9.3	16.3	40.4
		dirt-road T	39.1	85.0	29.1	32.1	46.0
		runway V	47.3	89.2	46.8	33.7	60.8
		runway T	46.2	87.7	43.6	43.5	48.8
4	HIT-UAV	dirt-road V	3.9	10.4	1.6	5.2	2.5
		dirt-road T	8.9	21.3	5.3	17.0	0.8
		runway V	28.7	65.5	19.5	27.6	29.9
		runway T	12.7	39.5	4.5	17.0	8.4
		HIT-UAV V	46.6	82.6	47.1	39.6	53.5
		HIT-UAV T	48.0	84.1	48.9	41.0	55.1
5	dirt-road +runway	HIT-UAV V	6.7	17.5	3.4	10.3	3.1
		HIT-UAV T	7.1	18.5	4.3	11.0	3.3
6	dirt-road	dirt-road V	22.0	67.5	4.3	15.2	28.9
		dirt-road T	32.7	79.4	19.6	33.0	32.4
		runway V	20.4	42.2	18.2	40.5	0.3
		runway T	18.0	50.7	9.0	19.2	16.8
7	runway	dirt-road V	15.2	44.5	3.8	1.5	29.0
		dirt-road T	21.3	58.7	5.3	15.0	27.7
		runway V	38.8	82.5	35.7	19.2	58.4
		runway T	46.2	87.5	43.0	45.6	46.8

5.2. Quantitative results

Transfer learning analysis. Tab. 3 reports different experiments of our transfer learning analysis that we obtained by using the Faster R-CNN detector [19]. We chose this detector because it is widely used in several benchmarks [2, 24]. We train models that are pre-trained on COCO [13].

In Exp. 1 we combine dirt-road and runway training data, and evaluate the performance on their respective validation and test splits. In addition to observing that the validation split is more challenging than the test split, the `person` category results to be the most difficult one to detect. This is mainly due to the background heat that makes the targets difficult to distinguish. In Exp. 2 we train without the bounding boxes that are flagged as occluded (still evaluating with the bounding boxes flagged occluded). We can observe

that the results are similar between the two experiments. This suggests that the use of bounding boxes flagged occluded appears to marginally help the `person` class, while slightly affecting the `vehicle` class. Because HIT-UAV contains both `person` and `vehicle`, in Exp. 3 we pre-train our detector on HIT-UAV [22] (in addition to starting from the model pre-trained on COCO) and then train with dirt-road+runway. HIT-UAV pre-training leads to improved performance compared to Exp. 1. In Exp. 4, we evaluate the transfer learning ability of the detector from HIT-UAV to both dirt-road and runway. Although HIT-UAV and MONET's categories and sensor modalities (thermal) are the same, experiments show that this is a rather challenging setting, i.e. the detector poorly generalises between these scenarios. Performances on runway are higher than those on dirt-road. This can be due to the fact that the structure of the environment of runway is more similar to that of HIT-UAV. We also report the upper bound on HIT-UAV in Exp. 4. In Exp. 5, we can see that there is poor transfer learning ability when training is on dirt-road+runway and evaluation is on HIT-UAV. Compared to HIT-UAV upper bound, we can observe that the performance gap in transfer learning is rather large. In Exp. 6 & 7 we report the transfer learning experiments focused on MONET's scenarios. Despite being recorded with the same sensor, training on one scenario and testing on the other leads to lower performances than the same scenario setting.

Detector comparisons. Tab. 4 reports the comparisons amongst the different detectors. Experiments were executed by including the bounding boxes flagged occluded and by zeroing the areas marked with the `ignore` label.

In the same-scenario setting, YOLOv8 and ObjectBox are the best performing ones. In the transfer learning setting, SSD and Def. DETR are the best performing ones, while YOLOv8 and VarifocalNet consistently perform second best. YOLOv8 consistently outperforms the other detectors in terms of AP_{75}, which indicates its superior ability in estimating the correct bounding box sizes. In terms of AP_{50}, results are mixed. Although the two settings were captured with the same camera and in similar locations, we can observe that object detection is generally rather challenging for all the detectors, especially in the transfer learning setting. All detectors perform poorly in dirt-road, in particular with the person category. Dirt-road is more challenging because it was captured during daytime, where ground heat makes human targets more difficult to distinguish than the vehicle targets. Except Def. DETR, all the detectors perform poorly on the vehicle class in dirt-road/runway. We believe that this is because the vehicles were captured from much fewer viewpoints in dirt-road than in runway, thus affecting generalisation when tested in runway. YOLOv8's original data augmentation strategies appears to be effective in dirt-road but less effective in runway, suggesting that a more tailored design of data augmentation for this problem could help improving the detection accuracy.

Table 4. Evaluation of several detectors initialised with COCO trained weights. Values are provided in percentage. Grey background indicates transfer learning results. Keys: F. R-CNN [19]: Faster R-CNN. Def. DETR [5]: Deformable DETR [5]. †: with data augmentations of the original paper. Bold indicates best, underline indicates second best.

Exp	Train/Eval	Detector	AP	AP_{50}	AP_{75}	class-AP person	class-AP vehicle
1	dirt-road/ dirt-road	F. R-CNN [19]	22.0	67.5	4.3	15.2	28.9
		SSD [15]	19.6	64.2	4.4	11.3	28.0
		CornerNet [11]	10.1	46.8	0.2	1.3	18.9
		FCOS [23]	14.7	55.2	0.4	3.1	26.3
		DETR [5]	12.5	44.1	0.7	0.7	24.3
		Def. DETR [5]	16.5	55.2	1.4	3.4	29.6
		VarifocalNet [27]	21.4	61.5	3.6	8.2	34.7
		ObjectBox [26]	**26.4**	**72.5**	5.7	**16.4**	36.5
		YOLOv8 [10]	25.1	64.9	**5.9**	11.0	**39.3**
		ObjectBox† [26]	31.4	68.1	**19.3**	15.4	**47.4**
		YOLOv8† [10]	**33.3**	**76.0**	15.6	**22.2**	44.4
2	runway/ dirt-road	F. R-CNN [19]	15.2	44.5	3.8	1.5	29.0
		SSD [15]	**21.9**	47.0	17.6	**1.6**	**42.1**
		CornerNet [11]	18.9	35.6	22.7	0.2	37.5
		FCOS [23]	19.2	47.0	10.8	0.3	38.0
		DETR [5]	8.8	26.9	0.7	0.0	17.6
		Def. DETR [5]	8.0	31.3	0.3	0.3	15.8
		VarifocalNet [27]	19.3	**47.1**	14.9	1.2	37.5
		ObjectBox [26]	14.9	36.9	4.8	1.0	28.8
		YOLOv8 [10]	20.8	34.0	**25.8**	0.5	41.0
		ObjectBox† [26]	**32.6**	62.9	28.4	**12.1**	53.0
		YOLOv8† [10]	**32.6**	**63.4**	**32.3**	11.4	**53.8**
3	runway/ runway	F. R-CNN [19]	38.8	82.5	35.7	19.2	58.4
		SSD [15]	42.6	85.1	40.9	25.2	60.0
		CornerNet [11]	39.6	76.4	37.1	20.2	58.9
		FCOS [23]	44.5	88.6	41.3	28.9	60.2
		DETR [5]	31.2	81.2	20.2	16.0	46.4
		Def. DETR [5]	44.1	**94.7**	39.9	28.8	59.5
		VarifocalNet [27]	49.5	92.4	46.3	37.1	61.9
		ObjectBox [26]	49.5	93.7	48.3	39.7	59.2
		YOLOv8 [10]	**53.5**	93.0	**58.8**	**42.0**	**65.1**
		ObjectBox† [26]	51.4	95.0	49.9	41.3	**61.6**
		YOLOv8† [10]	**52.4**	**95.8**	**54.9**	**45.0**	59.8
4	dirt-road/ runway	F. R-CNN [19]	20.4	42.2	18.2	**40.5**	0.3
		SSD [15]	18.6	43.1	13.6	37.0	0.3
		CornerNet [11]	22.7	51.9	18.0	20.2	25.2
		FCOS [23]	19.0	53.4	9.4	32.1	5.9
		DETR [5]	13.9	48.0	3.5	19.0	8.8
		Def. DETR [5]	**32.9**	**73.0**	26.0	33.8	**32.0**
		VarifocalNet [27]	30.7	**73.0**	21.8	35.3	26.1
		ObjectBox [26]	27.2	59.7	19.8	39.0	15.4
		YOLOv8 [10]	30.7	60.2	**28.3**	38.6	22.7
		ObjectBox† [26]	**32.4**	**63.2**	**29.7**	34.7	**30.1**
		YOLOv8† [10]	27.8	55.4	24.8	**34.9**	20.7

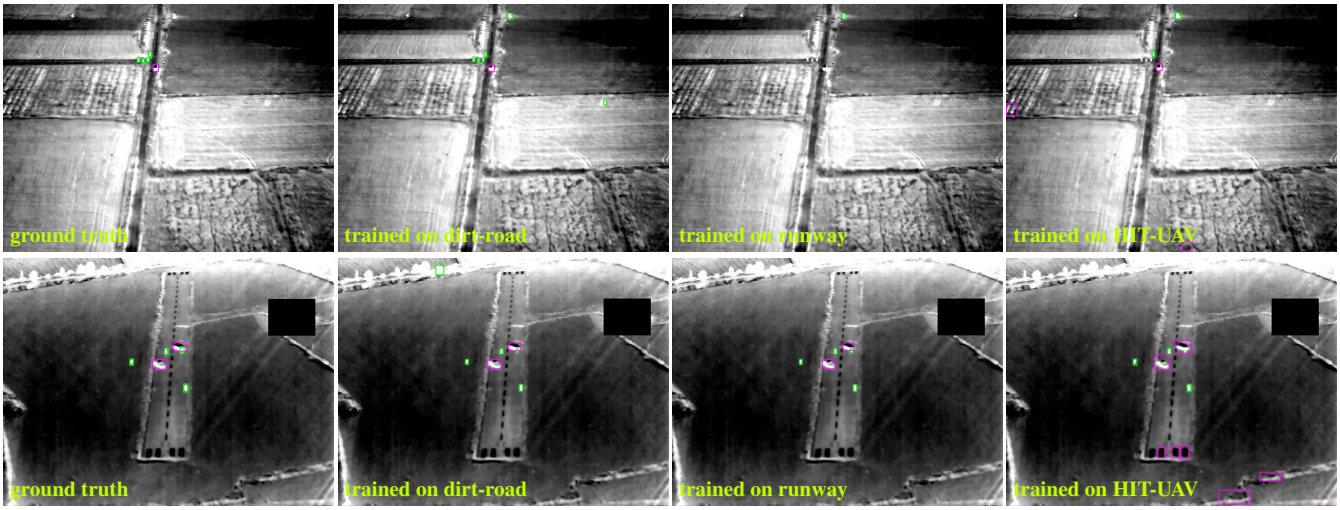


Figure 5. Qualitative results showing the performances of Faster R-CNN in the dirt-road (first row) and runway (second row) scenarios with different training datasets. Bounding boxes: green for `person`, magenta for `vehicle`.

Figure 6. Qualitative results showing the performances of Faster R-CNN in a scene from HIT-UAV with different training datasets.

5.3. Qualitative results

Fig. 5 shows examples of Faster R-CNN's detections on dirt-road (first row) and on runway (second row) produced with different training data. In the first case, training on dirt-road and evaluating on dirt-road leads to satisfactory results, when we train on runway the detector misses all the targets, and when we train on HIT-UAV the detector detects some targets plus some false positives. Conversely, in the second case, training on dirt-road and evaluating on runway leads to good results, same as when we train on runway, however when we train on HIT-UAV the detector produces several false positive detections. We observed that one of the reasons for this is that HIT-UAV contains several annotations of parked vehicles, which emits a low temperature, see example in Fig. 6. Hence, the detector relates dark patterns to vehicles, which are similar to the false positive detections in runway in Fig. 5. Fig. 6 also shows that training on MONET and evaluating on HIT-UAV leads to poor results. The Supplementary Material contains additional qualitative results from the nine detectors evaluated.

6. Conclusions

We introduce MONET, a novel and challenging multimodal dataset for vision-based object localisation in the thermal spectrum from drones. MONET was collected with a drone that flew over rural areas and captured human and vehicle activities. MONET comprises two scenarios in agricultural lands, namely *dirt-road* and *runway*. We benchmarked nine state-of-the-art object detection algorithms on MONET and found that they performed poorly due to the large scale variation of the targets, and the background clutter caused by the ground heat. The *dirt-road* scenario is more difficult than the *runway* scenario as it contains several more of the above-mentioned challenges than runway. To our knowledge, MONET is one of the few datasets in the thermal spectrum that provides a large number of manually annotated frames with timestamp-aligned metadata. Moreover, MONET includes bounding boxes with identities for each target, making it suitable for multi-object tracking research. We hope that MONET will foster further research, especially in the development of multimodal solutions for object localisation that exploits knowledge from metadata.

Limitations. Although we collected both RGB and thermal images and annotated both of them, for now we will not include the RGB images in this release of MONET due to still unaddressed privacy concerns. It is worth noting that only a portion of the dataset contains RGB images captured in light, while the others were captured in darkness.

References

[1] A. Berg, J. Ahlberg, and M. Felsberg. A thermal object tracking benchmark. In *AVSS*, 2015. 2

[2] E. Bondi, R. Jain, P. Aggrawal, S. Anand, R. Hannaford, A. Kapoor, J. Piavis, S. Shah, L. Joppa, B. Dilkina, and M. Tambe. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *WACV*, 2020. 2, 3, 6

[3] I. Bozcan and E. Kayacan. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *ICRA*, 2020. 2, 3

[4] C. Burke, P.R. McWhirter, J. Veitch-Michaelis, O. McAree, H.A.G. Pointon, S. Wich, and S. Longmore. Requirements and limitations of thermal drones for effective search and rescue in marine and coastal areas. *Drones*, 3(4), 2019. 1

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 6, 7

[6] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C.C. Loy, and D. Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv:1906.07155*, 2019. 6

[7] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), 2022. 4

[8] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *ECCV*, 2018. 2, 3

[9] M.R. Hsieh, Y.L. Lin, and W.H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *ICCV*, 2017. 2, 3

[10] G. Jocher, A. Chaurasia, and J. Qiu. YOLO by Ultralytics, 2023. 6, 7

[11] H. Law and J. Deng. CornerNet: Detecting objects as paired keypoints. In *ECCV*, 2018. 6, 7

[12] S. Li and D.-Y. Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI*, 2017. 2

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Dollar P., and C.L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 6

[14] Q. Liu, Z. He, X. Li, , and Y. Zheng. PTB-TIR: A Thermal Infrared Pedestrian Tracking Benchmark. 2019. 2

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 6, 7

[16] Y. Ma, X. Wu, G. Yu, and Y. Xu, Yand Wang. Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. *Sensors*, 16(4), 2016. 1

[17] M. Messmer, B. Kiefer, and A. Zell. Gaining scale invariance in UAV bird's eye view object detection by adaptive resizing. In *ICPR*, 2022. 2

[18] J. Portmann, S. Lynen, M. Chli, and R. Siegwart. People detection and tracking from aerial thermal views. In *ICRA*, 2014. 2

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 6, 7

[20] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016. 2, 3

[21] Stanford Artificial Intelligence Laboratory et al. Robotic Operating System, 2021. 4

[22] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi. HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicles. *arXiv:2204.03245*, 2022. 2, 3, 6, 7

[23] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 6, 7

[24] L.A. Varga, B. Kiefer, Martin Messmer, and Andreas Zell. SeaDronesSee: A maritime benchmark for detecting humans in open water. In *WACV*, 2022. 2, 3, 5, 6

[25] Z. Wu, N. Fuller, D. Theriault, and M. Betke. A thermal infrared video benchmark for visual analysis. In *CVPRW*, 2014. 1, 2

[26] M. Zand, A. Etemad, and M. Greenspan. Objectbox: From centers to boxes for anchor-free object detection. 2022. 6, 7

[27] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf. VarifocalNet: An IoU-aware dense object detector. *CVPR*, 2021. 6, 7

[28] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling. Detection and tracking meet drones challenge. *TPAMI*, 44(11), 2022. 2, 3

[29] X. Zhu, W. Su, L. Lu, B Li, X Wang, and J Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 6