scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection

Jiashun Suo[1,2], Tianyi Wang[3], Xingzhou Zhang[3] ✉, Haiyang Chen[1,2], Wei Zhou[1,2] & Weisong Shi[4]

**We present the HIT-UAV dataset, a high-altitude infrared thermal dataset for object detection applications on Unmanned Aerial Vehicles (UAVs). The dataset comprises 2,898 infrared thermal images extracted from 43,470 frames in hundreds of videos captured by UAVs in various scenarios, such as schools, parking lots, roads, and playgrounds. Moreover, the HIT-UAV provides essential flight data for each image, including flight altitude, camera perspective, date, and daylight intensity. For each image, we have manually annotated object instances with bounding boxes of two types (oriented and standard) to tackle the challenge of significant overlap of object instances in aerial images. To the best of our knowledge, the HIT-UAV is the first publicly available high-altitude UAV-based infrared thermal dataset for detecting persons and vehicles. We have trained and evaluated well-established object detection algorithms on the HIT-UAV. Our results demonstrate that the detection algorithms perform exceptionally well on the HIT-UAV compared to visual light datasets, since infrared thermal images do not contain significant irrelevant information about objects. We believe that the HIT-UAV will contribute to various UAV-based applications and researches. The dataset is freely available at https://pegasus.ac.cn.**

## Background & Summary

Unmanned Aerial Vehicle (UAV)-based object detection algorithms are widely used for various domains such as forest inventory[1], mapping applications[2], traffic monitoring[3], and humanitarian relief[4]. With the rapid development of deep learning[5] and edge computing[6], UAVs can now load edge computing devices to run artificial intelligence (AI) algorithms, thereby increasing their value in the aforementioned applications. Motivated by the rapid development of object detection, several general datasets such as PASCAL VOC[7], MSCOCO[8], and ImageNet[9] have been proposed to support algorithm training and evaluation. However, unlike natural environments, aerial images contain more object instances due to the wider view, bringing more significant challenges. Table 1 shows the average quantity of object bounding boxes per image for general datasets and the HIT-UAV[10]. Compared to general datasets, the HIT-UAV[10] contains a higher average quantity of object bounding boxes. Figure 1a,b use samples from the COCO and VisDrone datasets to show the differences between natural and aerial images.

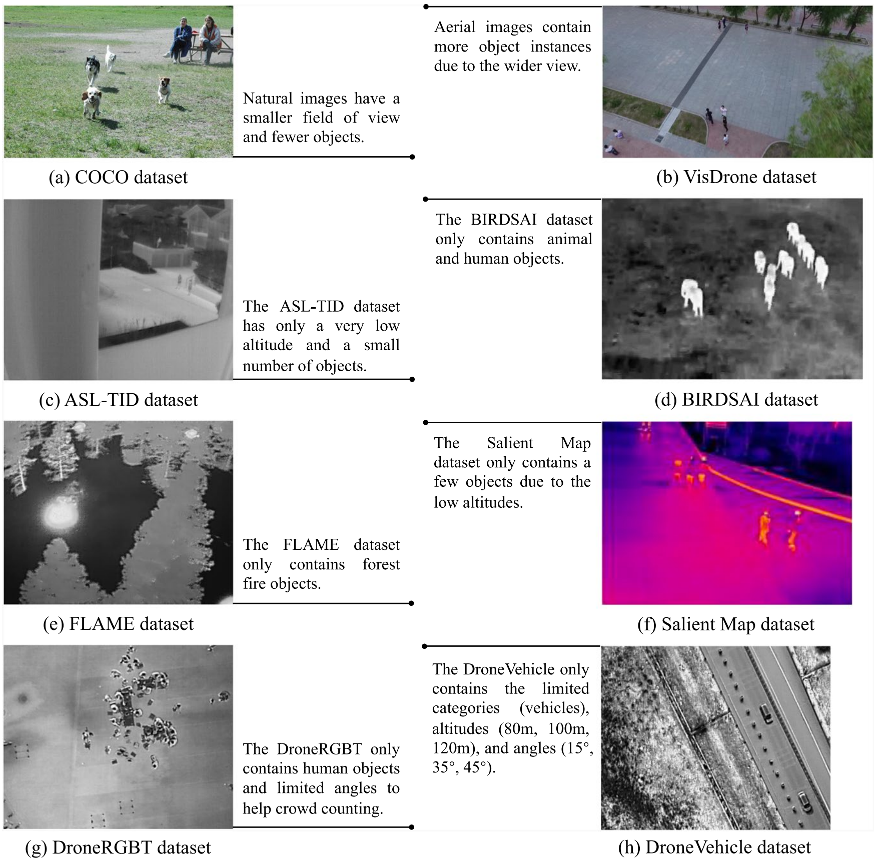
Many datasets of aerial perspectives have been introduced to help improve the detection performance of algorithms. The Stanford[11], UAV123[12], CARPK[13], VisDrone[14], and AU-AIR[15] datasets were introduced with visual light images. The ASL-TID[16], BIRDSAI[17], FLAME[18], DroneRGBT[19], DroneVehicle[20], and Salient Map[21] datasets were introduced with thermal infrared images. The Salient Map dataset contains pedestrian and vehicle objects because the authors found there is no publicly available thermal dataset for detecting pedestrians and vehicles from the perspective of UAVs.

However, although many datasets have been introduced for object detection on UAVs, there are many challenges in this field:

[1]Engineering Research Center of Cyberspace, Yunnan University, Kunming, 650091, China. [2]School of Software, Yunnan University, Kunming, 650091, China. [3]Research Center of Distributed Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China. [4]Department of Computer and Information Sciences, University of Delaware, Newark, DE, 19716, USA. ✉e-mail: zhangxingzhou@ict.ac.cn

| Dataset | Avg. Bbox |
|---|---|
| PASCAL VOC (2007 + 2012 version) | 2.89 |
| MSCOCO (2014 training + validation set) | 7.19 |
| ImageNet (2017 training set) | 1.37 |
| **HIT-UAV** | **8.59** |

**Table 1.** The average bounding box (Avg. Bbox) quantity of general datasets and the HIT-UAV.


Natural images have a smaller field of view and fewer objects.

(a) COCO dataset

Aerial images contain more object instances due to the wider view.

(b) VisDrone dataset

The ASL-TID dataset has only a very low altitude and a small number of objects.

(c) ASL-TID dataset

The BIRDSAI dataset only contains animal and human objects.

(d) BIRDSAI dataset

The FLAME dataset only contains forest fire objects.

(e) FLAME dataset

The Salient Map dataset only contains a few objects due to the low altitudes.

(f) Salient Map dataset

The DroneRGBT only contains human objects and limited angles to help crowd counting.

(g) DroneRGBT dataset

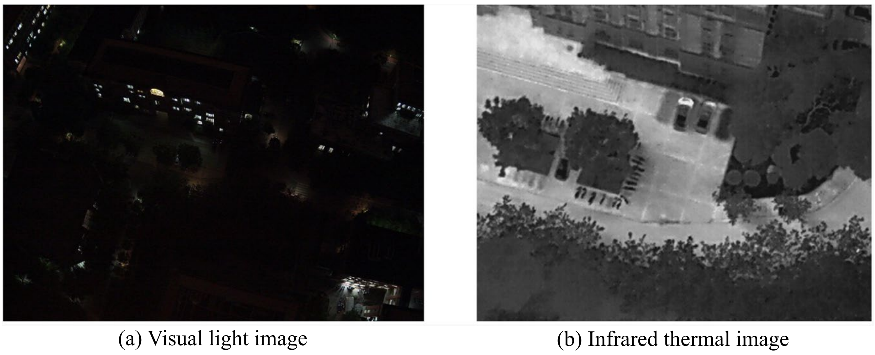
The DroneVehicle only contains the limited categories (vehicles), altitudes (80m, 100m, 120m), and angles (15°, 35°, 45°).

(h) DroneVehicle dataset

**Fig. 1** The samples of different datasets.

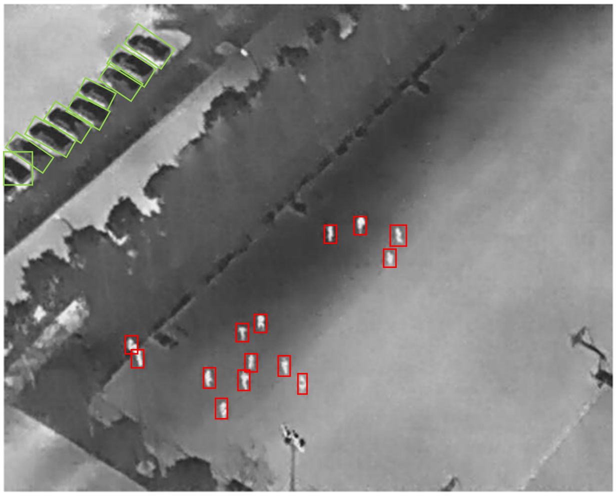
- *Limited application range*. Several extant UAV-based datasets only comprise visual light images, which limits their use during night-time operations and raises privacy concerns. As shown in Fig. 2, infrared thermal cameras offer distinct advantages over visual light cameras for night-time imaging. Additionally, Fig. 3 shows a sample image from the HIT-UAV[10], wherein persons are represented as white blocks devoid of any personal appearance, clothing, or gender information, thus ensuring complete protection of individual privacy.
- *Insufficient record information*. Numerous UAV-based datasets lack critical flight information, such as altitude and camera perspective, thereby precluding researchers from investigating pertinent issues, such as the influence of these factors on detection accuracy. Table 2 shows the record information of different datasets.
- *Non-diversified data distribution*. Many UAV-based datasets focus on a narrow range of aspects, such as synthetic scenes[12,17], low altitudes[12,13,16,21], single scenes[11,16], or specific object categories[13,18–20]. The limitations of synthetic scenes and low altitudes are highlighted in Fig. 4, which illustrates their drawbacks using sample images. Moreover, focusing on a single scene or object category restricts the applicability of the datasets in various scenarios, such as object detection in multiple scenes and detecting multiple object categories. To provide a comprehensive understanding of the current UAV-based infrared thermal datasets and their drawbacks, Fig. 1c–h are presented.

To overcome the aforementioned challenges, we present the HIT-UAV[10] dataset. The HIT-UAV[10] comprises infrared thermal images collected to expand the *application range* of UAVs at night. To facilitate research on diverse issues, such as the impact of UAV flight altitude and camera perspective on object detection accuracy, the HIT-UAV[10] records crucial *information*, including flight altitude, camera perspective, daylight intensity, and image shooting date. Figure 3 shows a sample image and the recorded information of the HIT-UAV[10]. Covering a wide range of aspects, including higher altitudes (ranging from 60 to 130 meters), different camera perspectives (ranging from 30 to 90 degrees),


(a) Visual light image

(b) Infrared thermal image

**Fig. 2** The sample images captured by visual light and infrared thermal cameras under the same flight altitude and camera angle at night. The infrared thermal image readily identifies car and bicycle objects, while the visual light image faces difficulty in doing so. The results demonstrate the superior performance of infrared thermal cameras in enabling UAVs to perform tasks more effectively during nighttime operations.



**Time**: night

**Date**: 15.01.2021

**Altitude**: 100m

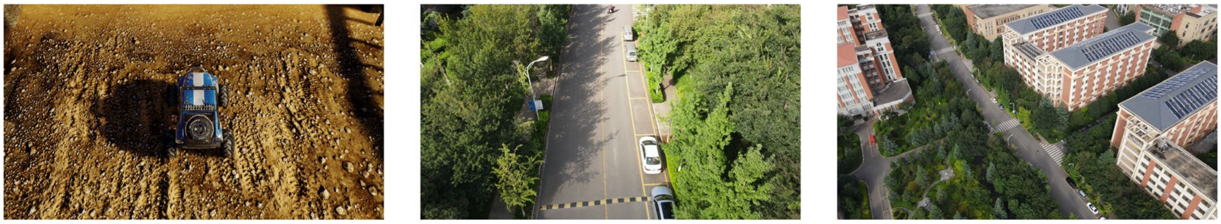
**Camera perspective**: 50°

**Weather**: no rain

**Fig. 3** A sample image and recorded information of the HIT-UAV.

| Dataset | Data type | Object annotation | Visual data | Altitude | Camera perspective | Infrared thermal |
|---|---|---|---|---|---|---|
| Stanford[11] | real | yes | yes | no | no | no |
| UAV123[12] | synthetic/real | yes | yes | no | no | no |
| CARPK[13] | real | yes | yes | no | no | no |
| VisDrone[14] | real | yes | yes | no | no | no |
| AU-AIR[15] | real | yes | yes | yes | no | no |
| ASL-TID[16] | real | yes | yes | no | no | yes |
| BIRDSAI[17] | synthetic/real | yes | yes | no | no | yes |
| FLAME[18] | real | no | yes | no | no | yes |
| DroneRBGT[19] | real | yes | yes | no | no | yes |
| DroneVehicle[20] | real | yes | yes | yes | yes | yes |
| Salient Map[21] | real | yes | yes | no | no | yes |
| **HIT-UAV**[10] | **real** | **yes** | **yes** | **yes** | **yes** | **yes** |

**Table 2.** The record information of different datasets.



(a) Synthetic scene of the UAV123 dataset

(b) Low-altitude real scene

(c) High-altitude real scene

**Fig. 4** The sample images from synthetic scenes, low-altitude real scenes, and high-altitude real scenes. Synthetic scenes often lack the lighting variations and details present in real scenes, which can result in poorer detection performance when models trained on synthetic scenes are applied to real scenes. Compared to low-altitude perspectives, high-altitude perspectives can detect more objects and enable UAVs to scan a larger area. Additionally, flying at higher altitudes allows UAVs to access areas with tall buildings, making high-altitude datasets advantageous for practical tasks. These advantages highlight the importance of high-altitude datasets in expanding the application of UAVs in real-world scenarios.

| | |
|---|---|
| **Dimensions** | Unfolded, 883 × 886 × 398 mm; Folded, 722 × 282 × 242 mm |
| **Diagonal Wheelbase** | 643 mm |
| **Weight** | Approx. 4.8 kg (with two TB55 batteries) |
| **Max Takeoff Weight** | 6.14 kg |
| **Max Payload** | 1.34 kg |
| **Max Angular Velocity** | Pitch: 300°/s, Yaw: 120°/s |
| **Max Ascent Speed** | 16.4 ft/s (5 m/s) |
| **Max Descent Speed (vertical)** | 9.8 ft/s (3 m/s) |
| **Max Speed** | S-mode/A-mode: 73.8 kph (45.9 mph); P-mode: 61.2 kph (38 mph) |
| **Max Flight Time (with two TB55 batteries)** | 34 min (no payload); 24 min (takeoff weight: 6.14 kg) |

**Table 3.** DJI Matrice M210 setup.

various scenes (such as schools, parking lots, roads, and playgrounds), and different common object categories (such as persons, cars, bicycles, and vehicles), the HIT-UAV[10] aims to increase *data distribution* for various tasks.

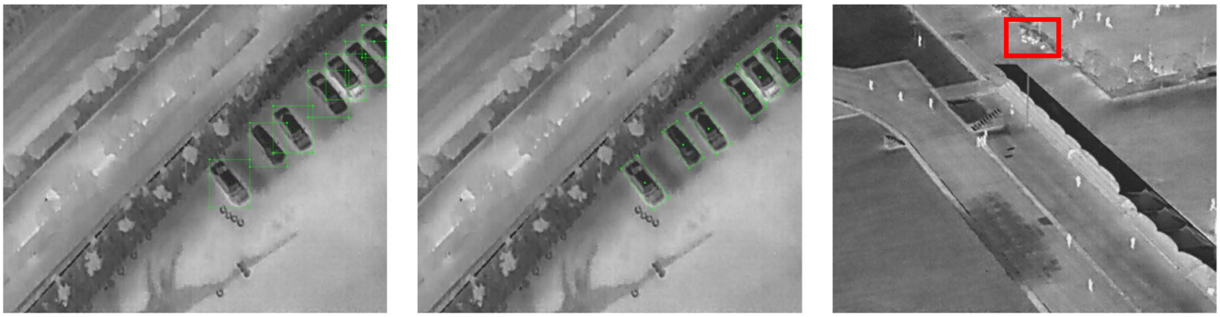
The dataset comprises 2,898 infrared thermal images extracted from 43,470 frames in hundreds of videos, and all frames were collected in public and desensitized. To promote effective use of the dataset on different tasks, the HIT-UAV[10] provides two types of annotated bounding boxes for each object in the images: oriented and standard. The oriented bounding box solves the issue of significant overlap between object instances in aerial images, while the standard bounding box facilitates efficient use of the dataset. The HIT-UAV[10] includes five object categories, namely *Person*, *Car*, *Bicycle*, *OtherVehicle*, and *DontCare*, with a total of 24,899 annotated objects. The *DontCare* category includes objects that could not be accurately categorized by the annotators (as further detailed in the Methods section). The dataset comprises 2,029 training images, 579 test images, and 290 validation images. To evaluate the HIT-UAV[10], we trained and tested the well-established object detection algorithms, namely YOLOv4[22], YOLOv4-tiny, Faster R-CNN[23], and SSD[24], using the dataset. The results show that compared to other visual light datasets, the algorithms exhibit exceptional performance on the HIT-UAV[10], indicating the potential of infrared thermal datasets to improve object detection applications in UAVs significantly. Further, we conducted an analysis of the performance of YOLOv4 and YOLOv4-tiny at different altitudes and camera perspectives, yielding insightful observations to aid users in their understanding of UAV-based object detection.

**To the best of our knowledge, the HIT-UAV[10] is the first publicly available high-altitude UAV-based infrared thermal dataset for detecting persons and vehicles. The HIT-UAV[10] has the great potential to enable several research activities**, such as (1) the application range of infrared thermal cameras in object detection tasks, (2) the feasibility of UAV-based search and rescue missions at night, (3) the relationship of flight altitude and object detection precision on UAVs, (4) the impact of camera perspective for UAV-based object detection.

## Methods

The UAV platform selected for image capture was the DJI Matrice M210 V2[25], which costs approximately 10,000 US dollars. The setup of the DJI Matrice M210 V2 used is detailed in Table 3. The DJI Zenmuse XT2 camera[26] was loaded on the UAV to capture the images. The DJI Zenmuse XT2 camera features a FLIR longwave infrared thermal camera with a thermal infrared camera resolution of 640 × 512 pixels and a 25 mm lens, as well as a visual camera that captures 4 K videos and 12MP photos. The cost of the DJI Zenmuse XT2 camera is approximately 8000 US dollars.

The dataset generation pipeline comprise four stages: video capture, frame extraction and data cleaning, object annotation, and dataset generation.



(a) Standard bounding box sample

(b) Oriented bounding box sample

(c) *DontCare* object sample

**Fig. 5** The samples of the standard bounding box, oriented bounding box, and *DontCare* object. Oriented bounding boxes have a smaller overlap than standard bounding boxes. In the (c), the red box represents the *DontCare* object. It is difficult to accurately identify whether the objects in this area are people or not.

**Video capture.** We captured videos under varying conditions, including schools, parking lots, roads, playgrounds, and more. The flight altitude ranged from 60 to 130 meters, and the camera perspective ranged from 30 to 90 degrees. We conducted flights during both day and night time. For each video, we recorded the flight altitude, camera perspective, flight date, and daylight intensity.

**Frame extraction and data cleaning.** There is a slight variation in image features between consecutive video frames, making most frames unsuitable for improving the performance of object detection model. Although many datasets reserve full frames to train detection models, this approach does not address the limited feature distribution problem. Fortunately, the HIT-UAV[10] provides a sufficient number of original frames (43,470 frames) to ensure a wide distribution of features. The frame resolution is 640 × 512, bit depth is 8, and the average compression rate is 21.059%. To filter adjacent frames that have little difference, we sampled an image every 15 frames (since the video refresh rate is 7 FPS), resulting in 2,898 infrared thermal images.

**Object annotation.** We annotated the objects in the dataset using two types of bounding boxes: standard and oriented. The standard bounding box is represented as $(x_c, y_c, w, h)$, where $(x_c, y_c)$ denotes the center coordinate and $w$ and $h$ denote the width and height of the bounding box, respectively. However, accurately labeling objects in aerial images from the perspective of UAVs can be challenging. To address this issue, we used $\theta$-based oriented bounding box[27] to label object instances. The oriented bounding box is represented as $(x_c, y_c, w, h, \theta)$, where $\theta$ denotes the oriented angle from the horizontal direction of the standard bounding box. As shown in Fig. 5a, the overlap of standard bounding boxes can be significant, making it difficult for state-of-the-art object detection algorithms to distinguish them well. Using oriented bounding boxes accurately annotates the objects and solves this issue, as shown in Fig. 5b. Note that the bounding box on the boundary is standard because the oriented bounding box cannot exceed the edge. One drawback of oriented bounding boxes is that few native object detection algorithms support training with them. To help users utilize the dataset, we provide both oriented and standard bounding box annotation files.

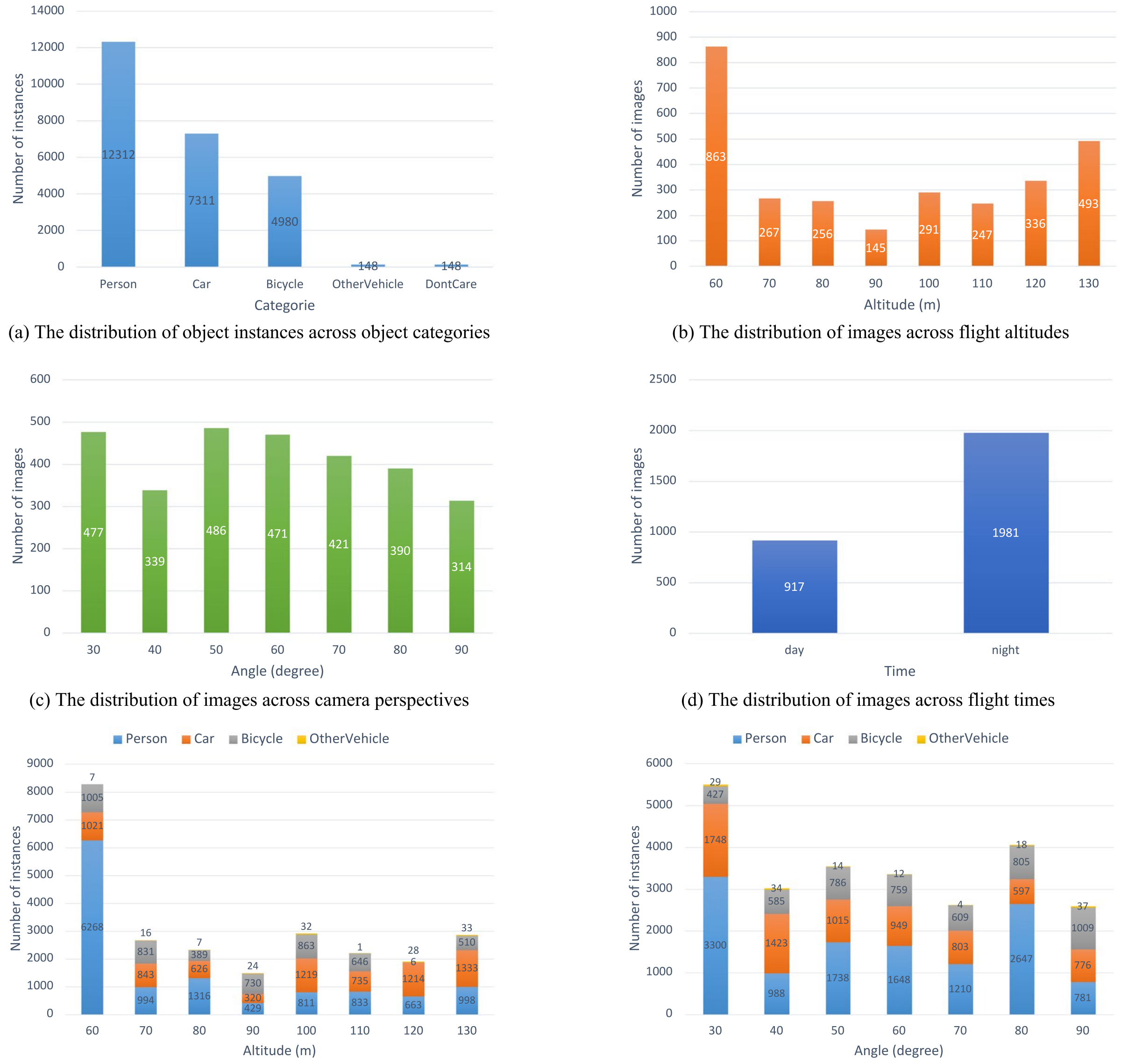
We performed manual annotation of oriented object bounding boxes for all images using a modified version of the LabelImg tool. Difficult and truncated object instances were also labeled. Three individuals were involved in the annotation process, and each annotation was verified by the others. To facilitate the use of the dataset, we developed a tool to convert oriented bounding boxes to standard bounding boxes. The conversion method is as follows: First, we obtained the minimum and maximum x and y coordinates $(x_{min}, x_{max}, y_{min}, y_{max})$ of the oriented bounding box. Then, we used $(x_{min}, x_{max}, y_{min}, y_{max})$ as the boundary to obtain the standard bounding box, where the center coordinate was calculated as $x_c = (x_{min} + x_{max})/2$ and $y_c = (y_{min} + y_{max})/2$, and the width and height were calculated as $w = x_{max} - x_{min}$ and $h = y_{max} - y_{min}$, respectively.

**Dataset generation.** We developed a dataset generation tool with functions that include XML and JSON label file generation and dataset splitting. The original images were organized into different folders based on flight data, and the tool generated XML and JSON label files corresponding to each image. To facilitate object detection model training, we split the dataset into training, test, and validation sets with a ratio of 70%, 20%, and 10%, respectively, using the Hold-out method[28].

## Data Records

The dataset is available at Zenodo[10].

**Folder structure and recording format.** We offer two types of annotation files for users: XML files based on the VOC dataset format and JSON files based on the MS COCO dataset format. Both of these formats are commonly used benchmarks for object detection in computer vision. The top-level folder of our dataset includes four subfolders: *normal_json*, *normal_xml*, *rotate_json*, and *rotate_xml*. The *normal_json* and *normal_xml* folders contain annotation files with standard bounding boxes in JSON and XML formats, respectively. On the other hand, the *rotate_json* and *rotate_xml* folders contain annotation files with oriented bounding boxes in JSON and XML formats, respectively.



(a) The distribution of object instances across object categories

(b) The distribution of images across flight altitudes

(c) The distribution of images across camera perspectives

(d) The distribution of images across flight times

(e) The distribution of object instances with different categories across flight altitudes

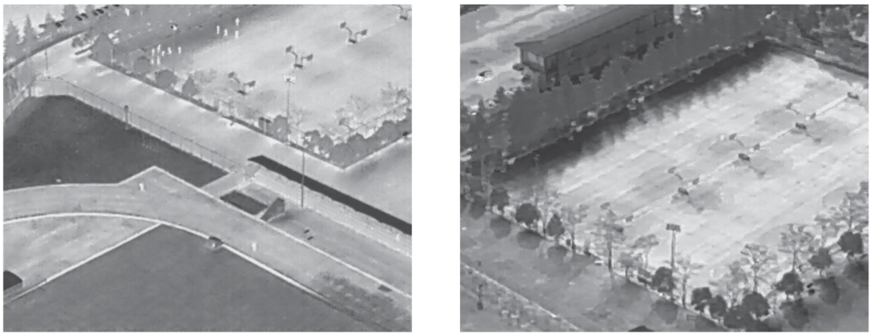
(f) The distribution of object instances with different categories across camera perspectives

**Fig. 6** The data distribution of the HIT-UAV.

The image files are named according to the following format: *T_HH*(*H*)*_AA_W_NNNNN*, where *T* indicates the shooting time (0 for day, 1 for night), *HH*(*H*) indicates the flight altitude (ranging from 60 to 130 meters), *AA* denotes the camera perspective (ranging from 30 to 90 degrees), *W* indicates the weather condition (only images captured under no rain conditions were included in the dataset), and *NNNNN* denotes the serial number of the image.

**Properties.** The annotated object categories include four types that highly appear in rescue and search missions: *Person*, *Car*, *Bicycle*, *OtherVehicle*. In addition, we labeled unrecognizable objects, namely *DontCare*, because many objects cannot identify specific types by annotator in high aerial images. As shown in Fig. 5c, the red box represents the object of *DontCare*. In this object area, it is difficult to accurately identify if they are persons. Therefore, the *DontCare* can point out easily confused objects in the image.

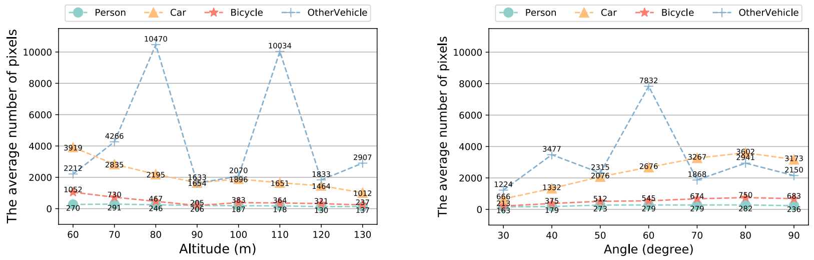
Figure 6a shows the distribution of annotations across object categories. The main object for the rescue mission (*Person*) appears more than other objects. Additionally, the presence of a substantial number of *Car* and *Bicycle* objects makes the HIT-UAV[10] suitable for a wide range of common tasks. To enhance the versatility of the dataset for high-altitude missions, flight altitudes were recorded in intervals of 10 meters, ranging from 60 to 130 meters. This information is depicted in Fig. 6b. The camera perspectives were also recorded in increments of 10 degrees, varying from 30 to 90 degrees, as shown in Fig. 6c. Infrared thermal images have a significant difference between day and night due to the higher background temperature during the day. As shown in Fig. 7, the infrared thermal image during the night is easier to identify the objects than during the day because the background temperature of the night is lower than the day. To increase the diversity of the dataset, infrared thermal images were collected both during the day and night, as presented in Fig. 6d. Figure 6e,f present the distribution of instances



(a) Night image sample

(b) Day image sample

**Fig. 7** The samples of the night and day images.



(a) The average pixel of categories across flight altitudes

(b) The average pixel of categories across camera perspectives

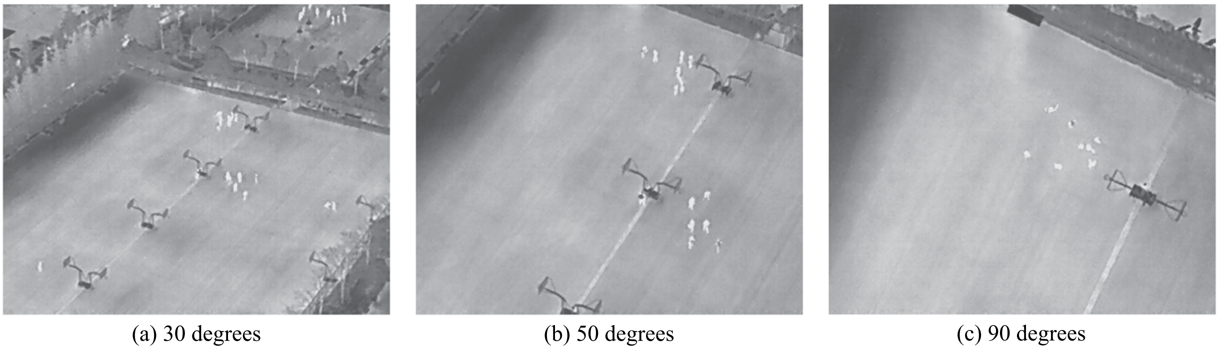
**Fig. 8** The average pixel of categories across flight altitudes and camera perspectives.

with varying categories across flight altitudes and camera perspectives, respectively. The average pixels of different categories across flight altitudes are depicted in Fig. 8a. Theoretically, the average pixel size is expected to decrease with increasing altitude, since higher altitudes result in smaller object sizes. However, for the category of *OtherVehicle*, the fluctuations are large due to its limited number of instances, which leads to the influence of truncated objects on the results. The average pixel of the remaining categories generally decreases with altitude, though there may be slight fluctuations due to the difference in image coverage at different altitudes and angles. The average pixel of categories across camera perspectives is shown in Fig. 8b, where it is observed that the average pixel size increases initially and then decreases with increasing angle. This is due to the fact that objects become more prominent with reduced vision, but their visible surface area decreases with greater angles. Figure 9 illustrates these visual changes.

## Technical Validation

We trained four well-established object detection algorithms, namely YOLOv4, YOLOv4-tiny, Faster-RCNN, and SSD, using the HIT-UAV[10]. The dataset consisted of 2,029 training images, 290 validation images, and 579 test images. The experiments were performed on an RTX 2080Ti GPU. YOLOv4 and YOLOv4-tiny were trained using the Darknet framework, while Faster-RCNN (with a ResNet-101 backbone) and SSD-512 were trained using the MMDetection[29] framework. The pre-trained models for YOLOv4 and YOLOv4-tiny were obtained from official sources. The training process was performed for a maximum of 10,000 steps, with a batch size of 64 and subdivision of 16. The learning rate was set to 0.0013 and was multiplied by 0.1 at steps 8000 and 9000. The weight decay and momentum were set to 0.949 and 0.0005. For Faster-RCNN and SSD, the official ResNet-101 and VGG16 models were used as pre-trained models. The maximum number of epochs was 32, with a batch size of 16. The learning rate was set to 0.02 and had a warm-up ratio of 0.001, with a warm-up iteration of 500. The weight decay and momentum were set to 0.9 and 0.0001.

Table 4 presents the precision of the aforementioned models on the HIT-UAV[10] test set, as well as the precision of YOLOv4 and YOLOv4-tiny trained on the COCO dataset and the highest accuracy (attained by RRNet) on the VisDrone-2019 challenge[30]. Our observations indicate that the Average Precision (AP) value for the category of *Person* is significantly lower when using YOLOv4-tiny on the HIT-UAV[10]. This discrepancy may be attributed to the lower detection capability of YOLOv4-tiny for small objects in comparison to other models. Additionally, the AP for the category of *OtherVehicle* is subpar, which may be due to the category imbalance issue. The SSD-512 model exhibits improved performance in the imbalanced category. In the VisDrone-2019 challenge, the highest precision of 55.82% mean Average Precision (mAP) was achieved by the RRNet method. However, the official YOLOv4 model achieved 65.7% mAP on the COCO dataset, surpassing RRNet in the VisDrone challenge. This indicates that aerial image information is more complex than that of natural images. Finally, for the HIT-UAV[10], YOLOv4 achieved an mAP of 84.75%, indicating the following observations:



(a) 30 degrees (b) 50 degrees (c) 90 degrees

**Fig. 9** The sample images taken at 80 meters with varying camera perspectives. At 30 degrees, objects in the far distance appear smaller due to the wider field of view. Conversely, at 50 degrees, objects appear larger. However, at 90 degrees, objects once again become smaller due to the reduction in the visible surface area of objects.

| Model | Dataset | *Person* AP (%) | *Car* AP (%) | *Bicycle* AP (%) | *OtherVehicle* AP (%) | mAP@0.50 (%) |
|---|---|---|---|---|---|---|
| YOLOv4 | HIT-UAV | 89.88 (TP = 2370, FP = 346) | 92.64 (TP = 1241, FP = 166) | 86.48 (TP = 696, FP = 158) | 69.99 (TP = 26, FP = 8) | 84.75 (TP = 4333, FP = 678, FN = 447) |
| YOLOv4-tiny | HIT-UAV | 16.86 (TP = 214, FP = 50) | 83.61 (TP = 1080, FP = 226) | 51.9 (TP = 398, FP = 182) | 49.17 (TP = 14, FP = 7) | 50.38 (TP = 1706, FP = 465, FN = 3074) |
| Faster-RCNN | HIT-UAV | 75.5 | 95.6 | 86.4 | 46.8 | 76.8 |
| SSD-512 | HIT-UAV | 85.6 | 96.3 | 86.0 | 74.4 | 85.6 |
| YOLOv4 | COCO | \ | \ | \ | \ | 65.7 |
| YOLOv4-tiny | COCO | \ | \ | \ | \ | 40.2 |
| RRNet | VisDrone-2019 | \ | \ | \ | \ | 55.82 |

**Table 4.** The Average Precision (AP) of the baseline models.

- Infrared thermal images effectively filter out extraneous information, leading to improved object identification.
- Infrared thermal images facilitate the outstanding performance of common detection models with limited image data, due to the easily recognizable features of the objects in such images. The HIT-UAV[10] has the potential to facilitate the detection of vehicles and persons by UAVs.

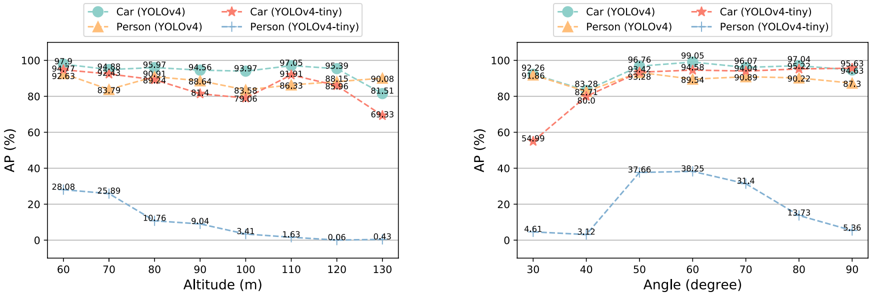
We used YOLOv4 and YOLOv4-tiny as samples to study the relationship and impact of altitude and camera perspective on UAV-based object detection. The categories of *Person* and *Car* were selected for this experiment, as the categories of *OtherVehicle* and *Bicycle* have a limited number of objects in the HIT-UAV[10]. A limited number of objects would result in fluctuations in statistical results. The results of the study are shown in Fig. 10. The following observations and insights have been gleaned from the results:

- The AP of YOLOv4 demonstrates stability within a certain range, suggesting that variations in altitudes and angles do not significantly impact the detection performance of robust algorithms.
- The AP of YOLOv4-tiny for the *Person* category tends to decrease with increasing altitude. This decrease is observed in three stages, ranging from 60 m to 80 m, 80 m to 90 m, and 100 m to 130 m, suggesting that the detection performance of lightweight algorithms is significantly impacted when objects fall outside of a certain size range. Higher altitudes provide a wider field of view, enabling UAVs to cover larger areas within the same flight time. In some UAV tasks, such as person rescue, users may need to weigh the trade-off between detection precision and altitude to achieve optimal performance.
- The AP of YOLOv4-tiny for the *Person* category first increases and then decreases with increasing camera angle. This result highlights the impact of the visible surface of objects on detection precision. At 90 degrees, as shown in Fig. 9c, individuals appear as points, making them more challenging to identify compared to when viewed at 50 degrees. As a result, it is crucial for users to choose the appropriate camera perspective when performing object detection tasks.

The sample detection results of the YOLOv4 model trained on the HIT-UAV[10] are shown in Fig. 11. The results demonstrate that the model effectively recognizes objects in infrared thermal aerial images. We hope the HIT-UAV[10] can promote the development of drone-based object detection tasks.

## Usage Notes

The HIT-UAV[10] is available at https://pegasus.ac.cn. Users can download the dataset to train object detection algorithms. The VOC and MS COCO dataset is a widely used benchmarks for object detection. We provide the label files with VOC and MS COCO format. Users can easily use the HIT-UAV[10].



(a) The AP of categories between different altitudes and models

(b) The AP of categories between different camera perspectives and models

**Fig. 10** The Average Precision (AP) of categories in the HIT-UAV test set.

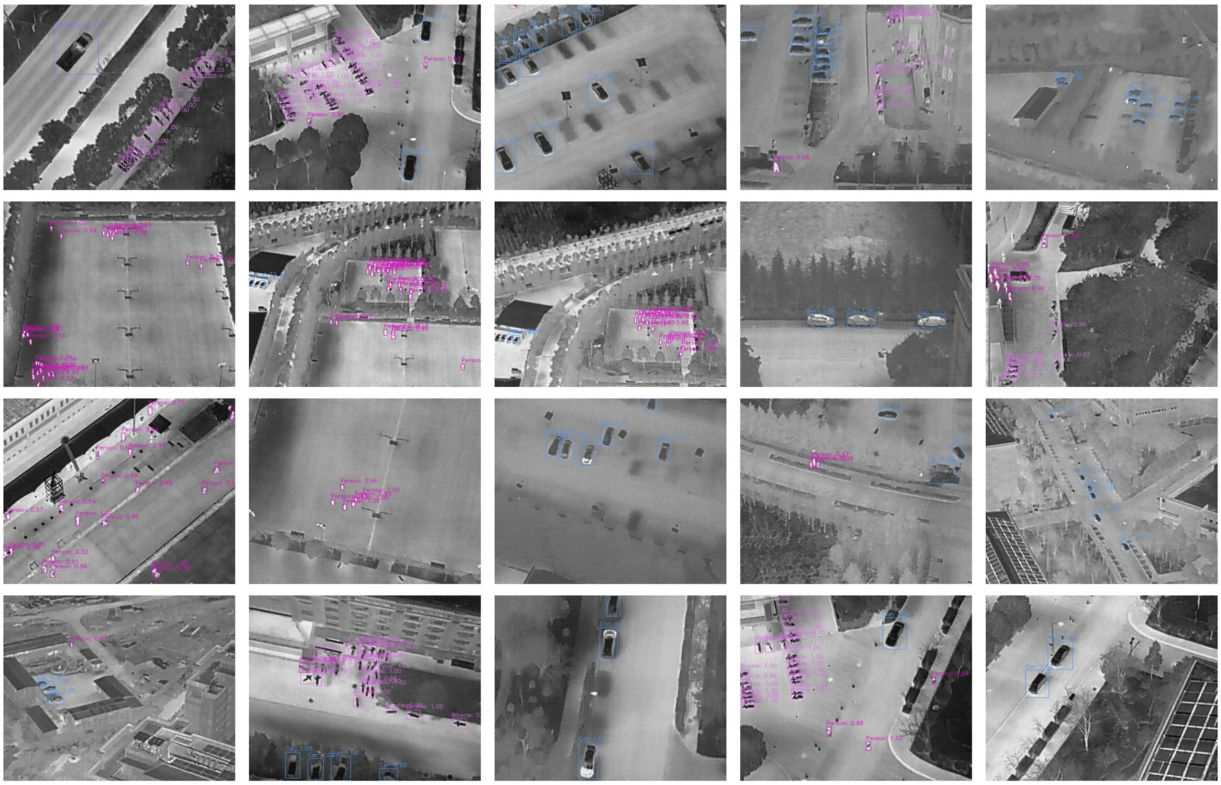


**Fig. 11** The sample results of YOLOv4 detection.

The HIT-UAV[10] was collected in a diverse range of environments, including schools, parking lots, roads, and playgrounds. This allows for the application of trained object detection models to these scenarios as well as other environments through the generalization capabilities of deep learning. Researchers can use the HIT-UAV[10] to train object detection models to research the application range of infrared thermal in different object detection tasks. Additionally, the trained models have the potential to be employed in UAV-based search and rescue missions during nighttime to evaluate their feasibility.

## Code availability

The data processing code is available in the *tools* folder of https://pegasus.ac.cn. The code is written in Python. The functions of the tools are as follows: (1) The *tools/devtoolkit/labelTransformer.py* is to convert oriented bounding boxes to standard bounding boxes and generate the dataset, (2) The *tools/devtoolkit/visualization.py* is to visualize images with bounding boxes, (3) The *tools/output/voc2yolo.py* is to generate the label files with the YOLO format to help users train the YOLO, which is the representative object detection algorithm.

Received: 28 September 2022; Accepted: 13 March 2023;
Published online: 20 April 2023

## References

1. Wallace, L., Lucieer, A., Watson, C. & Turner, D. Development of a UAV-LiDAR System with Application to Forest Inventory. *Remote Sens.* **4**(6), 1519–1543, https://doi.org/10.3390/rs4061519 (2012).
2. Samad, A. M., Kamarulzaman, N., Hamdani, M. A., Mastor, T. A. & Hashim, K. A. The potential of Unmanned Aerial Vehicle (UAV) for civilian and mapping application. *2013 IEEE 3rd International Conference on System Engineering and Technology* 313–318, https://doi.org/10.1109/ICSEngT.2013.6650191 (2013).
3. Heintz, F., Rudol, P. & Doherty, P. From images to traffic behavior - A UAV tracking and monitoring application. *2007 10th International Conference on Information Fusion* 1–8, https://doi.org/10.1109/ICIF.2007.4408103 (2007).
4. Bravo, R. Z. B., Leiras, A. & Cyrino Oliveira, F. L. The Use of UAVs in Humanitarian Relief: An Application of POMDP-Based Methodology for Finding Victims. *Production and Operations Management* **28**(2), 421–440, https://doi.org/10.1111/poms.12930 (2019).
5. Pouyanfar, S. *et al.* A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys (CSUR)* **51**(5), 1–36, https://doi.org/10.1145/3234150 (2018).
6. Shi, W., Cao, J., Zhang, Q., Li, Y. & Xu, L. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal* **3**(5), 637–646, https://doi.org/10.1109/JIOT.2016.2579198 (2016).
7. Everingham, M. *et al.* The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* **88**, 303–338 (2010).
8. Lin, T. *et al.* Microsoft COCO: Common Objects in Context. *Computer Vision–ECCV 2014: 13th European Conference* 745–755 (2014).
9. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (2009).
10. Suo, J. HIT-UAV: A High-altitude Infrared Thermal Dataset for Unmanned Aerial Vehicles (v1.2.1). *Zenodo* https://doi.org/10.5281/zenodo.7633134 (2023).
11. Robicquet, A., Sadeghian, A., Alahi, A. & Savarese, S. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. *Computer Vision-ECCV 2016: 14th European Conference* 549–565 (2016).
12. Mueller, M., Smith, N. & Ghanem, B. A Benchmark and Simulator for UAV Tracking. *Computer Vision–ECCV 2016: 14th European Conference* 445–461 (2016).
13. Hsieh, M., Lin, Y. & Hsu, W. H. Drone-Based Object Counting by Spatially Regularized Regional Proposal Network. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* 4145–4153 (2017).
14. Zhu, P., Wen, L., Bian, X., Ling, H. & Hu Q. Vision Meets Drones: A Challenge. Preprint at https://arxiv.org/abs/1804.07437 (2018).
15. Bozcan, I. & Kayacan, E. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. *2020 IEEE International Conference on Robotics and Automation (ICRA)* 8504–8510, https://doi.org/10.1109/ICRA40945.2020.9196845 (2020).
16. Portmann, J., Lynen, S., Chli, M. & Siegwart, R. People detection and tracking from aerial thermal views. *2014 IEEE International Conference on Robotics and Automation (ICRA)* 1794–1800, https://doi.org/10.1109/ICRA.2014.6907094 (2014).
17. Bondi, E. *et al.* BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 1747–1756 (2020).
18. Shamsoshoara, A. *et al.* Aerial Imagery Pile burn detection using Deep Learning: the FLAME dataset. *Computer Networks* **193**, 1008001, https://doi.org/10.1016/j.comnet.2021.108001 (2021).
19. Peng, T., Li, Q. & Zhu, P. RGB-T Crowd Counting from Drone: A Benchmark and MMCCN Network. *Proceedings of the Asian Conference on Computer Vision (ACCV)* (2020).
20. Sun, Y., Cao, B., Zhu, P. & Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(10), 6700–6713, https://doi.org/10.1109/TCSVT.2022.3168279 (2022).
21. Li, M., Zhao, X., Li, J. & Zhu, D. OBJECT DETECTION IN UAV-BORNE THERMAL IMAGES USING BOUNDARY-AWARE SALIENCY MAPS. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **43**, 1233–1238 (2020).
22. Bochkovskiy, A., Wang, C. Y. & Liao, H. Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. Preprint at https://arxiv.org/abs/2004.10934 (2020).
23. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems 28 (NIPS 2015)* **28**, (2015).
24. Liu, W. *et al.* SSD: Single Shot MultiBox Detector. *Computer Vision–ECCV 2016: 14th European Conference* 21–37 (2016)
25. DJI. *Matrice M210 V2* https://www.dji.com/matrice-200-series-v2 (2021).
26. DJI. *Zenmuse XT2* https://www.dji.com/zenmuse-xt2 (2021)
27. Yao, C., Bai, X., Liu, W., Ma, Y. & Tu, Z. Detecting texts of arbitrary orientations in natural images. *2012 IEEE Conference on Computer Vision and Pattern Recognition* 1083–1090, https://doi.org/10.1109/CVPR.2012.6247787 (2012).
28. Lee, L. C., Liong, C. Y. & Jemain, A. A. Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science. *Microchemical Journal* **139**, 125–133, https://doi.org/10.1016/j.microc.2018.02.009 (2018).
29. Chen, K. *et al.* MMDetection: Open mmlab detection toolbox and benchmark. Preprint at https://arxiv.org/abs/1906.07155 (2019).
30. Zhu, P. *et al.* Detection and Tracking Meet Drones Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 7380–7399, https://doi.org/10.1109/TPAMI.2021.3119563 (2021).

## Acknowledgements

This work was supported in part by the China Postdoctoral Science Foundation under Grant No.2021M693227, in part by the National Natural Science Foundation of China under Grant 62162067, 62101480, and 62072434, in part by the Yunnan Province Science Foundation under Grant No.202005AC160007, No.202001BB050076, Research and Application of Object Detection based on Artificial Intelligence. We are indebted to the anonymous referees for their constructive remarks, to the Cloud-Edge-Things Testbed of Nanjing Institute of Information Superbahn.

## Author contributions

Jiashun Suo designed the experiments, collected the data, annotated the images, and wrote the paper. Tianyi Wang annotated the images and coded the tools. Xingzhou Zhang reviewed and edited the paper and supervised the work. Haiyang Chen collected the data and annotated the images. Wei Zhou reviewed the paper, supervised the work, and provided funding. Weisong Shi reviewed and edited the paper. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to X.Z.

**Reprints and permissions information** is available at www.nature.com/reprints.