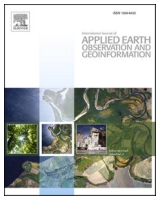
International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102912

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Object detection from UAV thermal infrared images and videos using YOLO models

Chenchen Jiang [a,b], Huazhong Ren [a,b,*], Xin Ye [a,b], Jinshun Zhu [a,b], Hui Zeng [a,b], Yang Nan [c], Min Sun [a,b], Xiang Ren [a,b], Hongtao Huo [d]

[a] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, China
[b] Beijing Key Laboratory of Spatial Information Integration and 3S Application, Peking University, Beijing 100871, China
[c] Shanghai Grandhonor Information Technology Co. Ltd, Shanghai 200333, China
[d] Graduate School, People's Public Security University of China, Beijing 100038, China

ARTICLE INFO

Keywords:
Object detection
Thermal infrared images and videos
YOLO

ABSTRACT

Object detection is one of the most crucial tasks in computer vision and remote sensing to identify specific categories of various objects in images. The unmanned aerial vehicle (UAV)-based thermal infrared (TIR) remote sensing multi-scenario images and videos are two important data sources in public security. However, their object detection process is still challenging because of the complicated scene information, coarse resolution compared with the visible videos and lack of public labelled datasets and training models. This study proposed a UAV TIR object detection framework for images and videos. The You Only Look Once (YOLO) models based on Convolutional Neural Network (CNN) architecture were designed to extract features from ground-based TIR images and videos, which were captured by Forward-looking Infrared (FLIR) cameras. The most effective algorithm was finally identified by evaluation metrics and then applied to detect objects on TIR videos from UAVs. Results showed that the highest mean average precision (mAP) of the person and car instances was 88.69% in the validating task. The fastest detection speed achieved 50 frames per second (FPS), and the smallest model size was observed in YOLOv5-s. In the application, the cross-detection performance on persons and cars in UAV TIR videos under a YOLOv5-s model was discussed in terms of the different UAVs' observation angles and the effectiveness of the YOLO architecture was revealed. This study provides positive support for the qualitative and quantitative evaluation of objection detection from TIR images and videos using deep-learning models.

1. Introduction

Object detection is a common task in computer vision and its main purpose is to classify and locate a specific object in an image. Remote sensing images contain multiple and multiscale objects captured by various sensors at different platforms and therefore provide promising and abundant data for object detection. By using various remote sensing platforms (Xiang et al., 2019), object detection has been performed on spaceborne, aerial, and ground remote sensing images. Ground remote sensing mainly refers to the remote sensing technology system with high towers, vehicles, and ships as platforms. These platforms are adopted to offer numerous optically labelled datasets but relatively few thermal infrared (TIR, with the wavelength range 8–14 μm) datasets for multiple objects at the ground level. UAV platforms can acquire high temperaspatial resolution images and videos, which makes up for the shortcomings of remote sensing satellites that are unavailability of high-resolution thermal images due to the limitations of satellite sensors. In addition, the rapid development of UAVs has led to increasing demand for highly efficient and effective intelligent detection algorithms. Currently, UAV TIR images are already available in precision agriculture (PA), which utilizes tools and technologies to identify variability in soil and crops in the field to improve farming practices and optimize agronomic inputs (Khanal et al., 2017). However, the UAV platforms equipped with TIR sensors have a massive amount of unlabelled data (as shown in Fig. 1), and they represent a new challenge in the development of transmission techniques, transfer algorithms, and detection applications from ground labelled data.

Compared with the optical sensors, TIR sensors can capture images in

* Corresponding author at: Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, China.
E-mail address: renhuazhong@pku.edu.cn (H. Ren).

https://doi.org/10.1016/j.jag.2022.102912
Received 26 October 2021; Received in revised form 22 June 2022; Accepted 7 July 2022
Available online 18 July 2022

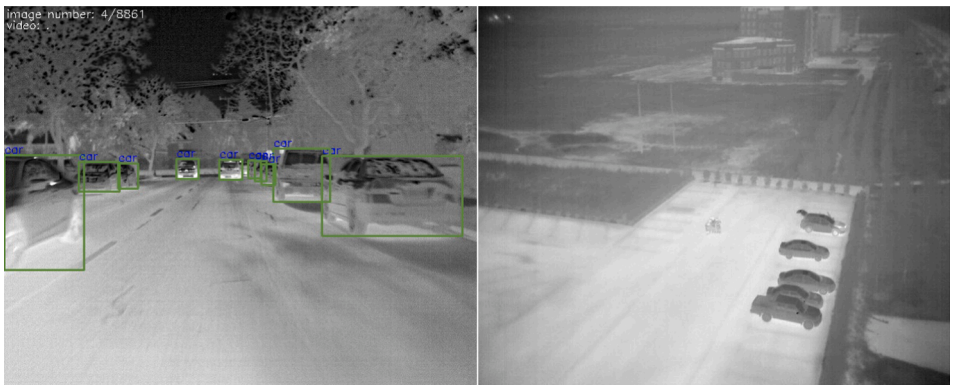


Fig. 1. A ground labelled thermal infrared image (left) versus a UAV unlabelled thermal infrared image (right).

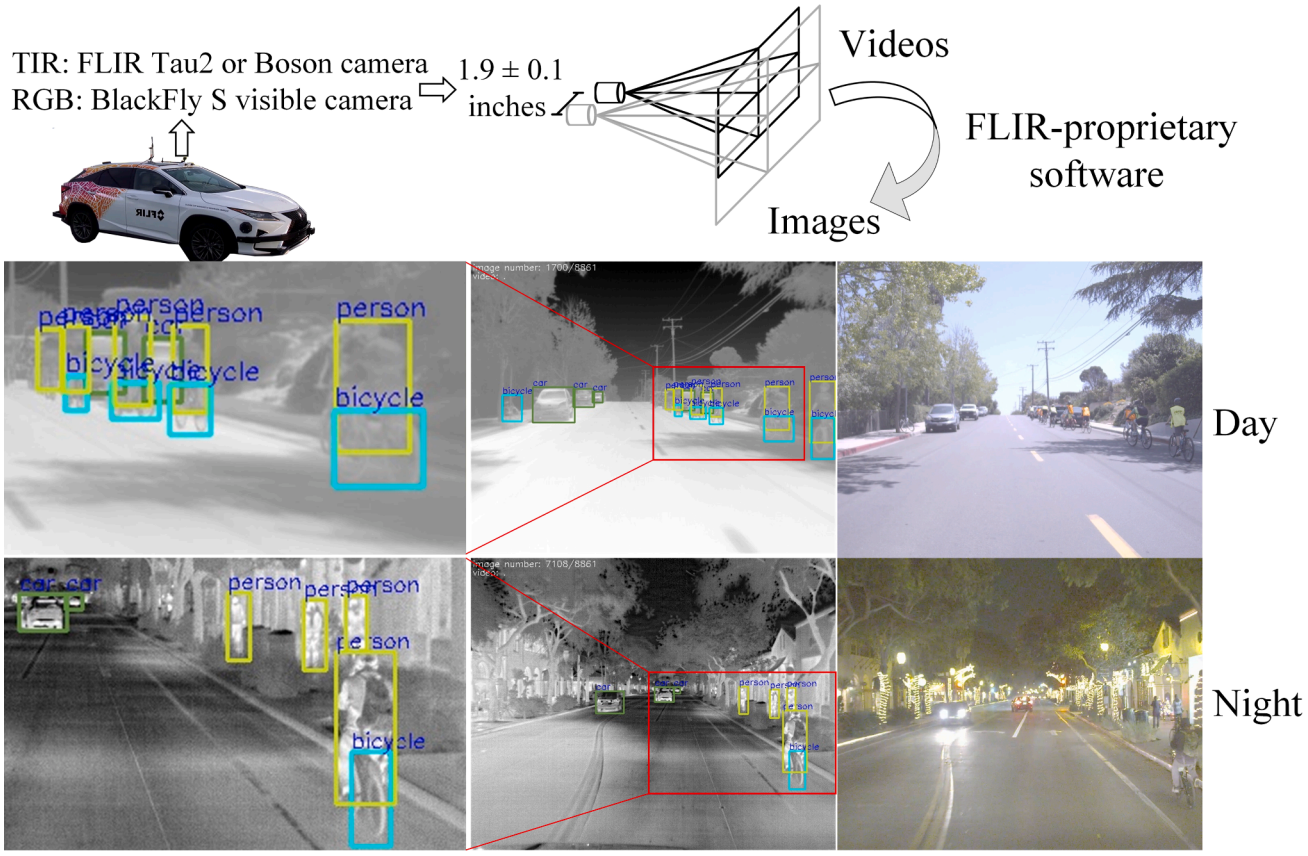


Fig. 2. FLIR dataset in RGB imagines and TIR imagines with annotations both Day and Night vision.

both day and night scenes. With the continuous improvement of TIR detection system performance, TIR technology has been widely used in body temperature detection, transportation surveillance system and public health and security fields and thus have received considerable attention. However, current object detection has not been extensively conducted on UAV TIR images and videos. Some efforts have been made on ground thermal infrared pedestrian detection (Chen and Shin, 2020; Kristo et al., 2020; Xu et al., 2019) by using computer vision and deep learning methods. Ships (Leira et al., 2015), vehicles (Zhang et al., 2018), thermal bridges in buildings (Garrido et al., 2018), and electrical equipment (Gong et al., 2018) in TIR images and videos, have also been studied. For example, Khalifa et al. (2019) surveyed different application and surveillance systems with embedded devices to detect human presence in different scenarios and compared the accuracy and performance time by using a unified dataset. Iwasaki et al. (2013) proposed a new method to detect vehicles by using TIR images and applied the method to monitor traffic flow automatically. They reported that the object detection method in thermal imagery could be effectively applied in road traffic monitoring.

Overall, object detection systems and UAVs are promising and growing technologies with many application scenarios not only in computer vision and deep learning but also in other areas, such as surveillance cases for the prevention and control of COVID-19, human detection in search and rescue, and Advanced Driver Assist Systems (ADAS) with thermal imaging technology (Kanistras et al., 2013; Rudol and Doherty, 2008; Shao et al., 2021). Although UAV TIR remote sensing has been applied in the above fields, object detection from UAV TIR images and videos still encounters numerous difficulties because of the complex image background, low resolution, long imaging distance, flight angles, and lack of public labelled datasets and TIR detection for

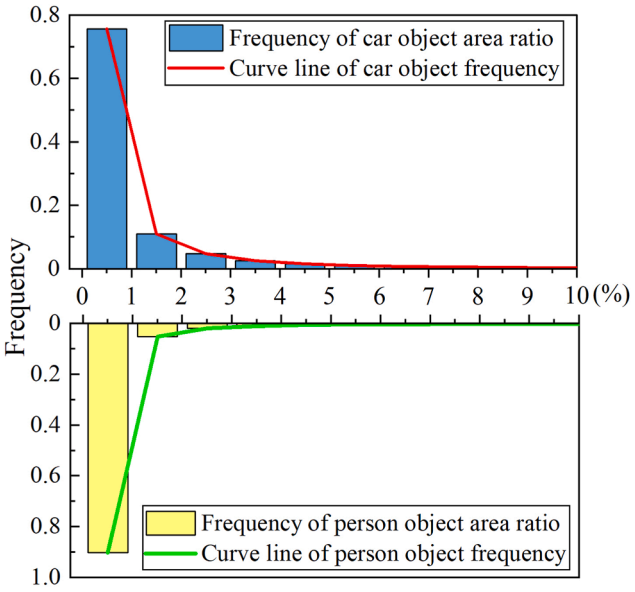


Fig. 3. Distribution of car and person instances scales.

Fig. 4. Multi-scenes of similar objects.

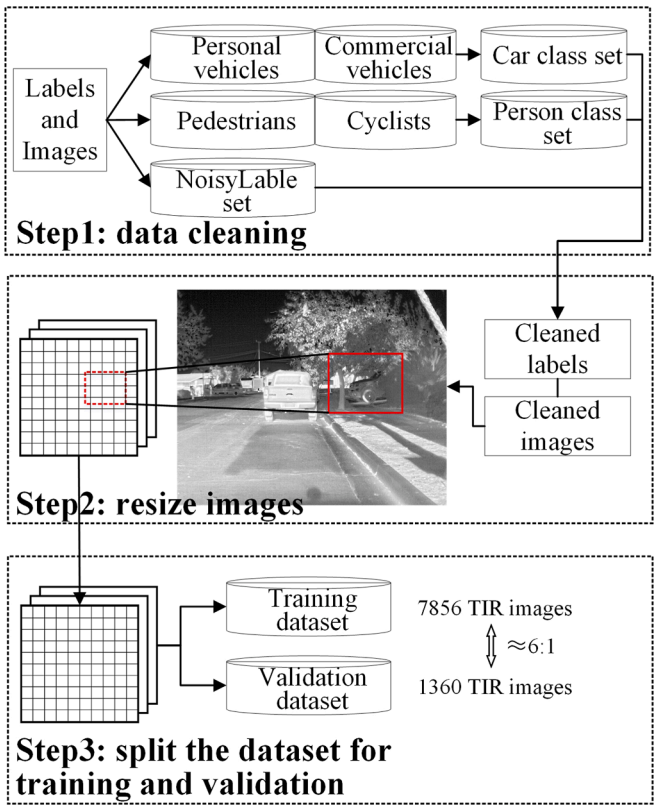


Fig. 5. Pre-processing of thermal infrared object labels and images.

multiple scenes and objects.

Deep learning is now widely used in image processing and object detection due to its powerful feature learning capabilities. At present, the most popular deep learning (DL)-based network architecture is the Convolutional Neural Network (CNN)-based architecture (Zhong et al., 2020). CNN-based object detection methods include two-stage detectors (e.g., Regions with CNN features (RCNN) (Girshick et al., 2014), Spatial Pyramid Pooling Networks (SPPNet) (He et al., 2015), Fast RCNN (Girshick, 2015), Faster RCNN (Ren et al., 2017) and Feature Pyramid Networks (FPN) (Lin et al., 2017a) and one-stage detectors (e.g., You Only Look Once (YOLO) (Redmon et al., 2016), Single Shot MultiBox Detector (SSD) (Liu et al., 2016) and RetinaNet (Lin et al., 2017b)). These successful detectors have been developed to detect and track objects in optical images (Li et al., 2020). Among them, the YOLO models with a one-stage detector have advantages in real-time optical image detection (Liu et al., 2020), but have relatively low accuracy for small objects. The two-stage detectors (R-CNN-based models) present high localization and accuracy, but the inference speed is relatively slow (Zou et al., 2019). Many UAV studies have tried to detect and track certain types of objects for autonomous navigation and landing in real-time (Hassan et al., 2019; Khoshboresh Masouleh and Shah-Hosseini, 2019; Lee et al., 2017; Zhang et al., 2019). The inference speed and model storage are obviously important for object detection applications of UAVs. In video detection, the NVIDIA-Jetson platform is always integrated with mobile devices to be able to perform online detection during movement (Ren et al., 2022; Shin and Kim, 2022), it has the restrictions of random access memory (RAM) and hard disk. One of the important trends in object detection is the shift to a faster and more efficient detection system. With the rapid development of one-stage object detection methods, the most popular and stable version of YOLO, which exhibits improved performance with multi-scale prediction boxes and a deep backbone network, was introduced by Redmon and Farhadi (2018). Bochkovskiy et al. (2020) presented YOLOv4 with several astounding new features, and the YOLOv4 outperforms YOLOv3 with a large margin in terms of accuracy and speed. Recently, Jocher et al. (2020) introduced YOLOv5 and a PyTorch-based version of YOLOv5 with exceptional improvements. The end-to-end network structure provides a detection rate that is higher than that of other networks (Kannadaguli, 2020).

To achieve unsupervised object detection on UAV TIR images and videos that lake labelled samples, this study focuses on evaluating different models, including YOLOv3, YOLOv4, and YOLOv5, for TIR multi-scenario and multi-object detection in bright and dark conditions for car and person instances. From this point of view, the network of YOLO series models on the ground TIR images dataset will be firstly retrained and examined, and then the trained YOLO models will be applied to detect objects from UAV TIR videos observed from different angles. As a result, this paper is organized as follows. Section 2 will show ground TIR images datasets and UAV TIR videos that will be used in network training and application; Section 3 will present technical approaches to object detection using YOLO models, and Section 4 will summarize and discuss the results of the YOLO models' object detection while Section 5 will focus on presenting a cross-application in UAV TIR videos. Finally, the paper will be ended with some conclusions and discussions for future research.

2. Dataset and processing

2.1. Ground TIR image dataset

The FLIR Thermal Starter Dataset (FLIR, 2019) was used and it provides four classes, namely, person, car, bicycle, and dog, with annotated thermal images and non-annotated optical images for training and validation of object detection neural networks. The dataset consists of 10,288 annotated TIR images with 28,151 persons and 46,692 cars, and 4224 video annotated TIR images with 21,965 persons and 14,013 cars. These images were captured on streets and highways in Santa Barbara, California, USA from November to May under generally clear sky conditions at day and night with a FLIR Tau2 camera (13 mm f/1.0, 45° horizontal field of view (HFOV) and 37° vertical field of view (VFOV)) or a Boson camera. Optical images were acquired with a FLIR Blackfly S camera (4–8 mm f/1.4–16 megapixel lens with the field of view (FOV) set to match Tau2). Both cameras were operated in default mode. The cameras were in a single enclosure 1.9 ± 0.1 in. apart from each other, as shown in Fig. 2.

Compared with conventional optical images, the TIR images have a lower spatial resolution, smaller signal-to-noise ratio (SNR), and fewer texture features. However, the TIR images and videos can be taken at day and night, while the optical images can only be collected in the daytime if no artificial light is provided. In order to combine the advantages of all-day observations of thermal infrared images and the features of optical images, we assumed that the shape features that YOLO models have learned from optical datasets (Lin et al., 2014) for the person or car are similar in TIR images, and thus provide a reasonable pre-trained model for detection in thermal infrared images (Kristo et al., 2020).

2.1.1. Scale and scenes

We compared object scales by calculating the area ratio of object bounding box pixels to the total image. The statistics on histogram and curve line distribution in the pixels of objects area were investigated respectively on car and person instances, as shown in Fig. 3. The majority of the area ratios for the observed car and person instances were below 10 %. Moreover, 1 % of the area ratio of the cars and persons accounted for 76 % and 90 %, respectively. The statistics on the bounding box area ratio of cars and persons showed that the dataset was dominated by medium-scale (cars) and small-scale (persons) instances, which are usually more important and difficult than large objects because at the level of applications (Tian et al., 2020; Zhao et al., 2018), it is necessary to generate an efficient framework with short reaction

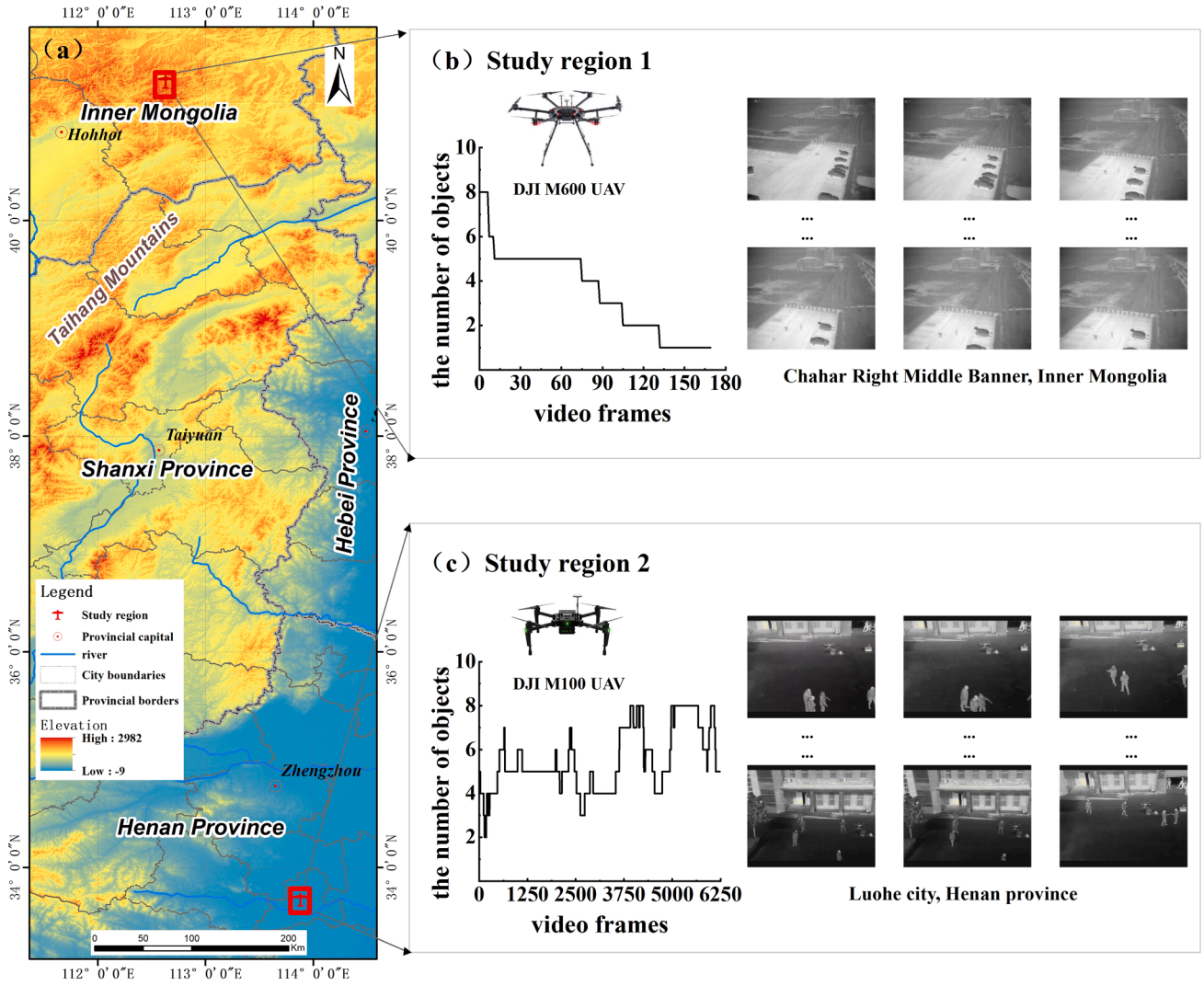


Fig. 6. A map of the study regions and UAVs TIR videos. (a) study regions; (b) and (c) two detailed maps of the UAV TIR dataset observed by different study regions (the number of objects and frames for each study region is shown in the line graphs of Fig. 6 (b) and Fig. 6 (c) respectively).

time to detect small-scale objects. Meanwhile, medium-scale and small-scale objects can effectively evaluate the accuracy of object detection algorithms.

The multi-scenario TIR images for the person and car instances are shown in Fig. 4. Multiple scenarios with similar cars or persons are presented in the four rows. The first two rows show that a single object class (pedestrians or vehicles) existed in multiple scenes. Streets and highways contained cars and persons in a single image, as seen on the third row of Fig. 4. Pedestrians and cyclists are shown in the second and fourth rows, respectively. Overall, we grouped the cars and persons in various scenes into four categories, namely, stopping cars, moving cars, pedestrians, and cyclists.

2.1.2. Data pre-processing

Fig. 5 shows the data pre-processing procedure adopted in this study. We traversed the objects of the whole dataset and defined personal vehicles and some small commercial vehicles as Car class, pedestrians and cyclists as Person class, and the other objects as noisy labels. Then, the cleaned thermal infrared images were resized to the same size for training and validation. It totally includes 9216 thermal infrared images, which are separated into training data (7856 images) and validation data (1360 images), following a ratio of nearly 6:1 (Liu et al., 2020).

2.2. UAV video datasets and processing

In this study, two experimental regions were selected in Luohe City, Henan province, and Chahar Right Middle Banner, Inner Mongolia, respectively. TIR sensors installed on two UAV platforms were used to collect TIR images and videos at both day and night. An overview of study regions and the UAV video datasets are provided in Fig. 6.

The datasets contain two individual UAV TIR datasets for cross object detection applications: a higher zenith angled dataset (study region 1 in Fig. 6 (a)) and a low flight height dataset (study region 2 in Fig. 6 (b)). The dataset was acquired by a higher zenith angle on the DJI M600 UAV platform in Chahar Right Middle Banner, Inner Mongolia at 23:00 on July 15, 2019. The DJI M600 UAV platform was equipped with DJI XT2 dual sensor, it acquired a 30-second TIR video, in which there are 1 to 8 vehicle objects in each frame. In addition, the low flight height dataset was acquired by the DJI M100 UAV platform in Luohe City, Henan Province at local 12:00 on December 17, 2018. The DJI M100 UAV platform acquired a 4-minute TIR video, and there are 6233 frames with person instances.

The flight characteristics of two UAV platforms of different objects in two datasets are presented in Fig. 7. The UAV platform on the left had a low flight height, and its flight path was translational. By contrast, the UAV on the right flew at a higher zenith angle, and the route was forward. These custom UAV TIR videos include static and dynamic objects

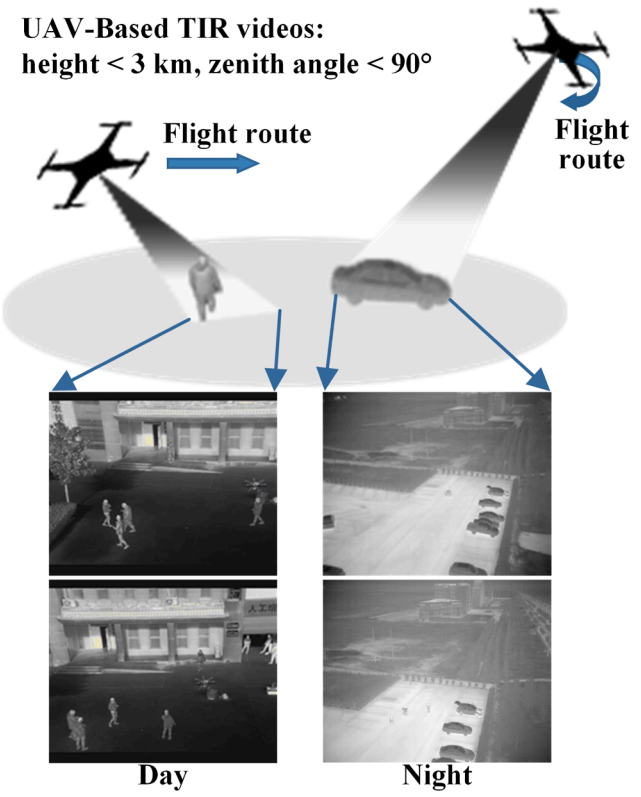


Fig. 7. UAVs TIR images from different viewing angles.

(persons and cars), which will be used in the following test task.

3. TIR object detection models and techniques

In this section, the YOLO framework for TIR object detection with ground TIR images and UAV-borne TIR videos is described. As shown in Fig. 8, the acquired three-band ground TIR images were divided into a grid. Convolutional layers are then used to extract and fuse the in-depth features of the objects from the input patch, and output the predicted multiple bounding boxes and class probabilities for those boxes. We assign Intersection Over Union (IoU), confidence score, and confusion matrix as evaluation metrics on the accuracy of predicted bounding boxes versus ground truth. To evaluate the detection speed, one of the most common metrics is frames per second (FPS), which is used in this study to evaluate the speed of TIR video and image object detection. In cross-application, we focused on the object detection of UAV-borne TIR videos under different observation angles. In summary, this framework can detect multiple objects in the ground TIR images and evaluate the precision and speed to achieve cross-application in UAV-borne TIR videos under different observation angles.

3.1. YOLO series models

YOLO series end-to-end real-time target detection methods have been proposed in recent years as the third versions (Redmon and Farhadi, 2018), the fourth versions (Bochkovskiy et al., 2020) and the fifth versions (Ultralytics, 2020a) of the YOLO detector. In this study, these YOLO models in Table 1 were compared and evaluated. Some studies showed that the YOLO method performs well in PASCAL VOC and COCO RGB image datasets (Huang et al., 2020; Zhao and Li, 2020). Consequently, a wide range of YOLO baseline-based novel methods have been proposed in recent years (Kumar et al., 2021; Tian et al., 2020).

Table 1 compared different YOLO versions from the five aspects of backbone, neck, activation, loss and model architectures. In the backbone aspect, the Darknet-53 network structure was used in the YOLOv3 model (Redmon and Farhadi, 2018), which consists primarily of a series of 1 × 1 and 3 × 3 convolutional layers, each of which is followed by a Batch Normalization (BN layer) and a LeakyReLU layer and has 53 convolutional layers (CBL). YOLOv4 utilized a Cross Stage Partial Network (CSPDarknet-53) and an open-source neural network framework as the main backbone network to train and extract image features (Wu et al., 2020). YOLOv5 was added with a focus layer designed for floating-point operations per second (FLOPS) reduction and speed increase. The Feature Pyramid Networks (FPN) and Path Aggregation Network structure (PANet) were employed as the neck network to achieve an improved fusion of the extracted features (Wang et al., 2019). The Spatial Pyramid Pooling (SPP) layer was used in the YOLOv4 model to transform the convolution features of different sizes into pooled features with the same length (He et al., 2015). Compared with YOLOv3, YOLOv4 and YOLOv5 adopt mosaic data enhancement in the data processing.

YOLOv3 and YOLOv5 use LeakyReLU as the activation function. The benefit of using LeakyReLU is that during backpropagation, the gradient can be calculated for the parts of the LeakyReLU activation function that have an input less than zero (instead of having a value of zero as with ReLU), thus avoiding the serration of the gradient directions. Moreover, YOLOv4 utilizes the Mish activation function in Backbone. Several studies have shown that Mish tends to match or improve the performance of neural network architectures as compared to ReLU and LeakyReLU (Misra, 2019).

The loss functions of YOLOv3, YOLOv4, and YOLOv5 are summarized into three parts: the error caused by (x, y, w, h) for various bounding box positions (the box regression loss); the loss caused by confidence (the object loss); and the classification loss (the class loss). Some calculation methods can be adopted for the bounding box loss, for example, the Complete IoU (CIoU), Generalized IoU (GIoU), and Distance IoU (DIoU) (Zheng et al., 2020).

Generally, the data processing and network architecture are optimized, which makes YOLOv4 fast and produces the optimal balance between accuracy and speed in real-time object detection algorithms (Bochkovskiy et al., 2020). Although the structure of YOLOv5 does not exhibit much improvement, the storage size of YOLOv5 is smaller than that of YOLOv4. YOLOv5 network has mosaic data augmentation, CSP and PANet key innovations compared with YOLOv3, which improves the performance of object detection, particularly small object detection (Chen et al., 2021).

3.2. Evaluation metrics

For the object detection training and validation task in this work, precision, recall, mAP (mean average precision), and FPS were utilized to evaluate and compare the performance of the YOLO models.

Precision is used to measure the proportion of true positive samples in all the predicted positive samples, and recall is used to measure the proportion of true positive samples in all the predicted positive samples. Eqs. (1) and (2) show the calculation for precision and recall:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

TP (True Positive) represents the number of cars or persons that are correctly recognized as cars (or persons), FP (False Positives) means the number of samples that identified non-car instances (or non-person instances) as cars (or persons), and FN (False Negatives) indicates the number of samples that identified cars (or persons) as non-car instances (or non-person instances).

To enhance box localization performance, the IoU (Intersection over Union) score between the predicted bounding box and the correspond-

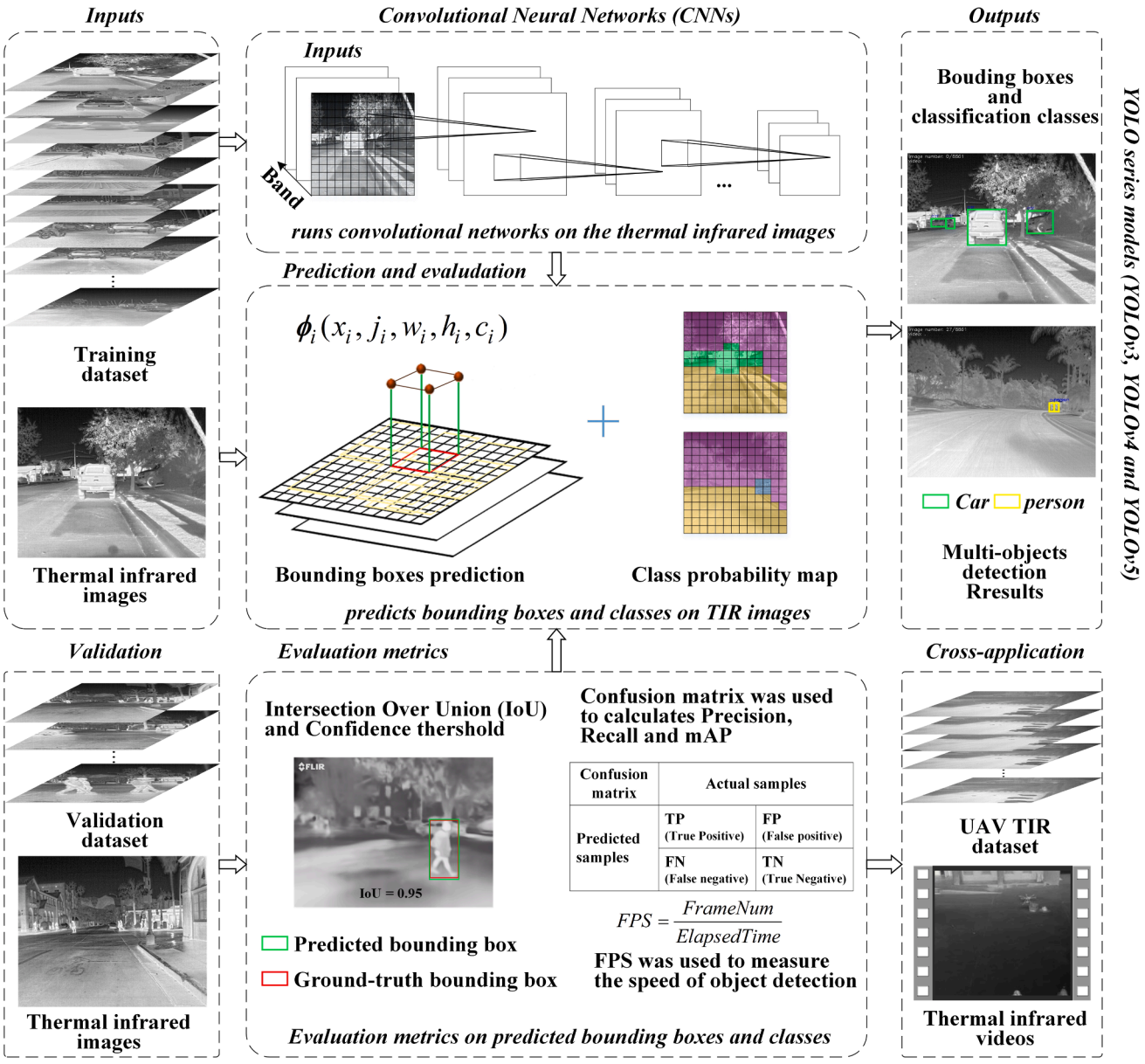


Fig. 8. Flowchart of the object detection framework in this study.

Table 1
Comparison of YOLOv3, YOLOv4 and YOLOv5 in network structure.

	YOLOv3	YOLOv4	YOLOv5
Backbone	Darknet-53	CSPDarknet-53	Focus, CSP
Neck	CBL	SPP, FPN, PAN	FPN, PAN
Activation	LeakyReLU	Mish	LeakyReLU
Loss		Bounding box, Class, Object	
Models	YOLOv3, YOLOv3-spp, YOLOv3-tiny	YOLOv4, YOLOv4-tiny, YOLOv4-pacsp, YOLOv4-pacsp-mish, YOLOv4-pacsp-s, YOLOv4-pacsp-s-mish, YOLOv4-pacsp-x, YOLOv4-pacsp-x-mish	YOLOv5-s, YOLOv5-m, YOLOv5-l, YOLOv5-x

ing annotation bounding box was considered. Mathematically, IoU can be described as Eqs. (3):

$$\mathrm{IoU} = \left|\frac{B_g \cap B_p}{B_g \cup B_p}\right| \tag{3}$$

Where B_g and B_p indicate ground-truth and predicted bounding boxes. If the IoU score exceeds a certain threshold, it is considered a true positive detection result. We calculate the mAP across different IoU thresholds, and the final metric mAP across test data is calculated by taking the mean of average precisions (AP) of all classes. The key parameters of testing detection are empirically set as confidence threshold = 0.25 and IoU = 0.65, and they are uniformly applied to all testing experiments. Fig. 9 shows an example of a positive and negative object detection bounding box considering the IoU score in the case of person

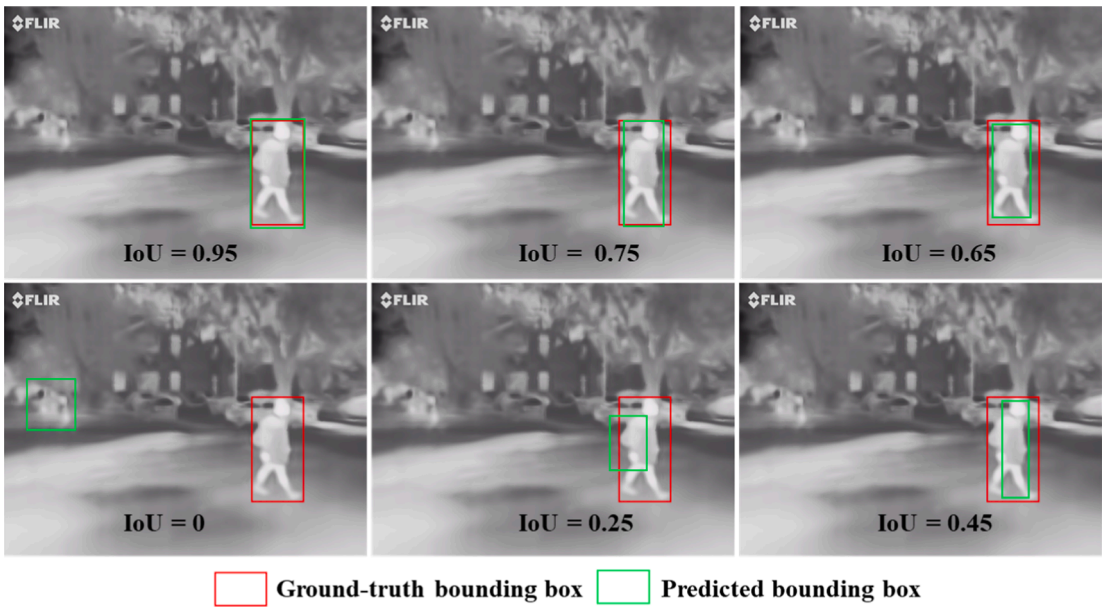


Fig. 9. Visual positive (top) and negative (bottom) representations of IoU with different thresholds.

detection in TIR images.

Additionally, for the detected infrared video, we extracted 833 images after every-four frames from the FLIR videos and 6402 frames from the UAV TIR videos, then used FPS as the speed performance metric in detecting the frames of the captured TIR videos.

3.3. Technical approaches of object detection using YOLO models

The multi-scenario object detection procedures of YOLOv3, YOLOv4, and YOLOv5 models for the FLIR dataset and UAV TIR video detection are shown in Fig. 10. The procedure contains several steps:

Step 1: Data pre-processing, as stated in section 2.

Step 2: Training processing with the pre-trained model configurations involved 15 models based on the COCO dataset, the pre-defined network, and the number of key parameters used. The software and hardware platforms adopted in the training experiment were implemented on a computer with a 2.5 GHz CPU, 8 GB RAM, GeForce GTX 1080ti GPU, CUDA10.2, and cuDNN7.0. YOLOv3, YOLOv4, and YOLOv5 were implemented by Python and PyTorch. The key parameters of Wang et al.'s (2020) and Ultralytics's (2020a, 2020b) programs (available in GitHub) were used in training the models. In consideration of the hardware, the mini-batch size was set to 8 due to the lack of memory in the GPU for algorithms with large convolutional layers and parameters. The maximum-batch size was set to 16. All models were trained from pre-trained checkpoints with the TIR images and their annotations, and the entire training process was terminated in 300 epochs on the training and validation datasets. We set the initial learning rate at 10^{-2} for the first epoch and adopted warm-up stage and cosine decay stage for scheduling the learning rate (Bochkovskiy et al., 2020; Loshchilov and Hutter, 2017). The momentum and weight decay were set as 0.9 and 0.005, respectively (Bochkovskiy et al., 2020). For data augmentation, all of our training experiments used the same hyperparameters as the default setting in the HSV color space. The weights and loss algorithms of the generators were initialized using YOLOv3, YOLOv4, and YOLOv5 models provided in section 3.1.

Step 3: During the testing stage, the feature of a testing video image dataset was used for testing the trained models. To evaluate the trained models under various architectures, and to test the optimal weights that can help determine the optimal YOLO model for application in Step 4, this study used the precision, mAP, and speed (FPS) metrics which were shown in section 3.2.

Step 4: An unlabeled UAV TIR video was searched using those different models that had been already tested on thermal video images from other sensors. All the frames of the UAV videos were detected in the Top1 aggregative indicator model to realize real-time automated person detection in UAV TIR videos.

4. Experiment results and validation

4.1. Model training

We conducted a preliminary experiment to evaluate the performance of state-of-the-art optical object detectors for car and person detection in a TIR dataset. The performance in multi-scenario object detection and the solution in YOLO was examined with the FLIR dataset to select the neural network architecture to be used in the testing part. We retrained YOLOv3, YOLOv4, and YOLOv5 without changing their original architecture on 7856 TIR images with 640 × 512 resolution from the FLIR dataset for 300 epochs and validated them on 1360 images from the FLIR validation set. In particular, mAP@0.5:0.95 was the average of different IoU settings from 0.5 to 0.95 with a step size of 0.05 for training. The performance and comparative results of YOLO are given in Fig. 11.

The precision and recall values in Fig. 11 (a) and (b) indicate that the overall recall scores in the various YOLO architectures were above 80 %, and the precision scores were between 45 % and 80 %. The training experiment showed that the precision scores of YOLOv3-spp and standard YOLOv3 (70.38 % and 70.31 %, respectively) were superior to the other models. The YOLOv3-spp model achieved the 88.91 %, which was the second-highest recall score. Similarly, the precision of standard YOLOv3 was the second-highest precision among all the models. The YOLOv3 experiment indicated that the standard YOLOv3 and YOLOv3-spp models could detect many car and person instances in the images. Focusing on the performance of the YOLOv4 models, we found that YOLOv4-pacsp-x model achieved a precision score of 62.64 % with a recall of 84.87 %, higher than those of YOLOv3 and YOLOv4 with simple architectures. This result means that YOLOv4-pacsp-x could detect more false positives in the images compared with the other models. For the YOLOv5 object detection training task, the models with ranked precisions from low to high are respectively YOLOv5-s, YOLOv5-m, YOLOv5-l, and YOLOv5-x with 60.60 %, 61.71 %, 62.26 %, and 63.62 %, respectively; with regard to the recall score, YOLOv5-l with 88.29 %

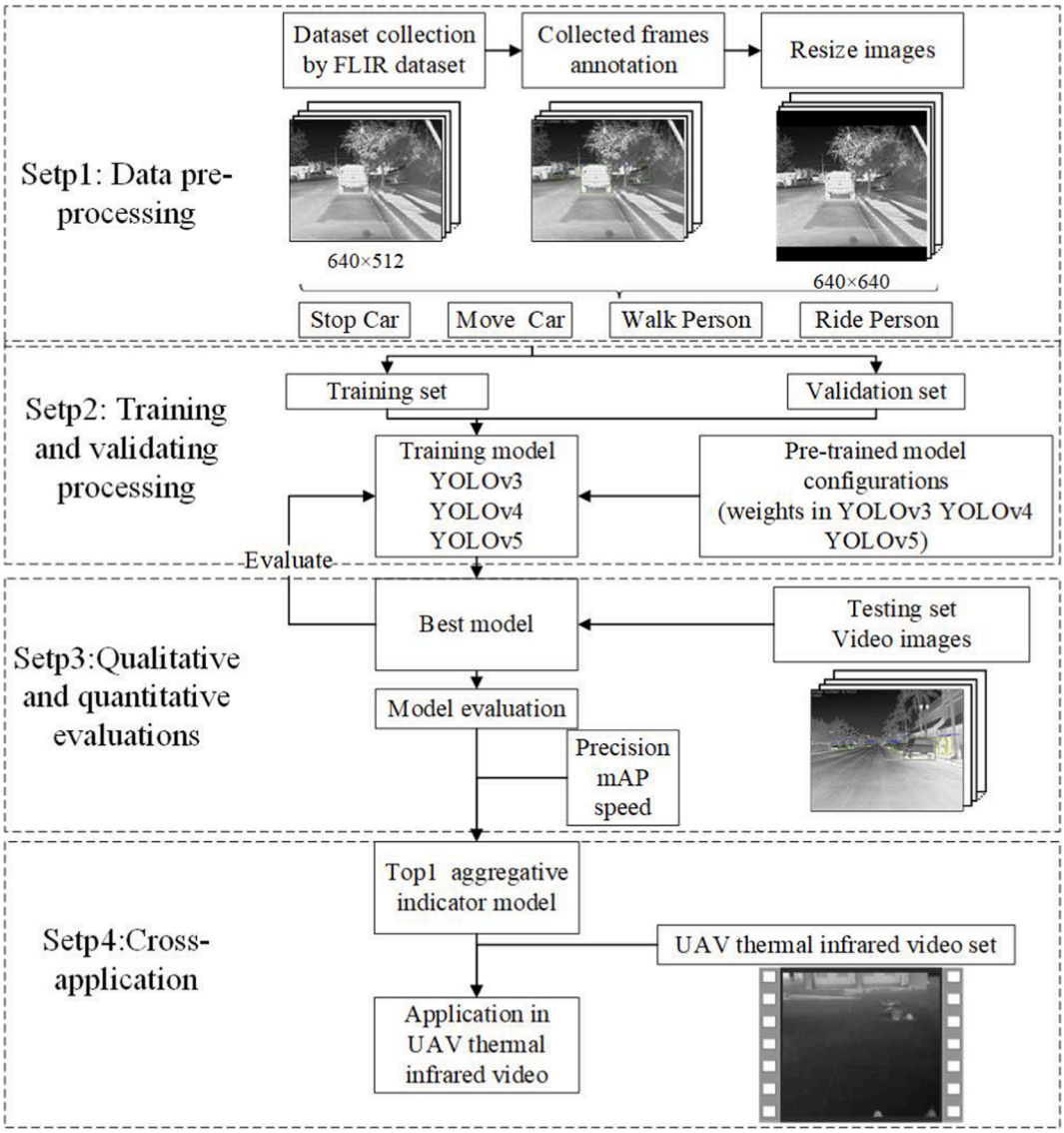


Fig. 10. The technical approach in multi-scenario thermal object detection using YOLO.

was better than the others.

Furthermore, we performed a mAP comparison on the validation training data, as shown in Fig. 11 (c) and (d). For the task of detecting two classes (i.e., car and person), the model with the highest mAP@0.5 validation was YOLOv3 at 88.69 %. YOLOv3 and YOLOv4 with corresponding networks showed better performance than these simple architectures with regards to car and person detection. However, the score of mAP@0.5 in the YOLOv5-s model was 86.75 %, which even outperformed that of standard YOLOv4. The mAP@0.5 scores of the YOLOv5 models were extremely close to 87 %, and the scores of mAP@0.5 in the YOLOv4 architectures ranged between 83 % and 87 %. However, compared with mAP@0.5:0.95, which is the average IoU set from 0.5 to 0.95 with a step size of 0.05, the most sophisticated stochastic approach in the score of mAP@0.5:0.95 was YOLOv5-x with 54.46 %. The small convolutional layers and parameters used in YOLOv3 and YOLOv4 (41.01 % and 40.10 %, respectively) have lower precision compared with the other models. The scores of mAP@0.5:0.95 of the standard YOLOv4 and a series of YOLOv4-pacsp were between 44.69 % and 48.84 %. The YOLOv5 model resulted in a mAP@0.5:0.95 score that was nearly 52.81 % compared with the majority of the models.

Fig. 11 (e) shows the loss curves of all training models. The loss of each model is mostly below 0.1 at 300 epochs and the difference is relatively small. YOLOv3 and YOLO3-spp have a relatively lower loss compared with the other models.

In conclusion, from the perspective of precision, recall, and mAP, the standard YOLOv3, YOLOv3-spp, YOLOv4-pacsp and YOLOv5-x can effectively train models with low missed-detection and false detection.

4.2. Validation and optimum model selection

For the object detection-testing task, precision and speed were also used to evaluate the performance of the training models. For the quantitative assessment of training and testing results, these approaches in YOLO deep learning architecture are summarized in Table 2 to evaluate their precision in multi-object cases, recall, mAP, FPS, and the computer storage. The test results for all classes showed that the mAP and precision of YOLOv3-spp, YOLOv4-pacsp, YOLOv4-pacsp-mish, and YOLOv5-s was much higher than that of the other models in their networks probably because the dataset is small (even smaller in the training network), and deep networks increase overfitting in the training process. Moreover, Table 2 indicates that YOLOv3 and YOLOv3-spp were superior to YOLOv3-tiny in detecting cars and persons in 300 epochs. YOLOv3-tiny exhibited better performance for a large car instance in the

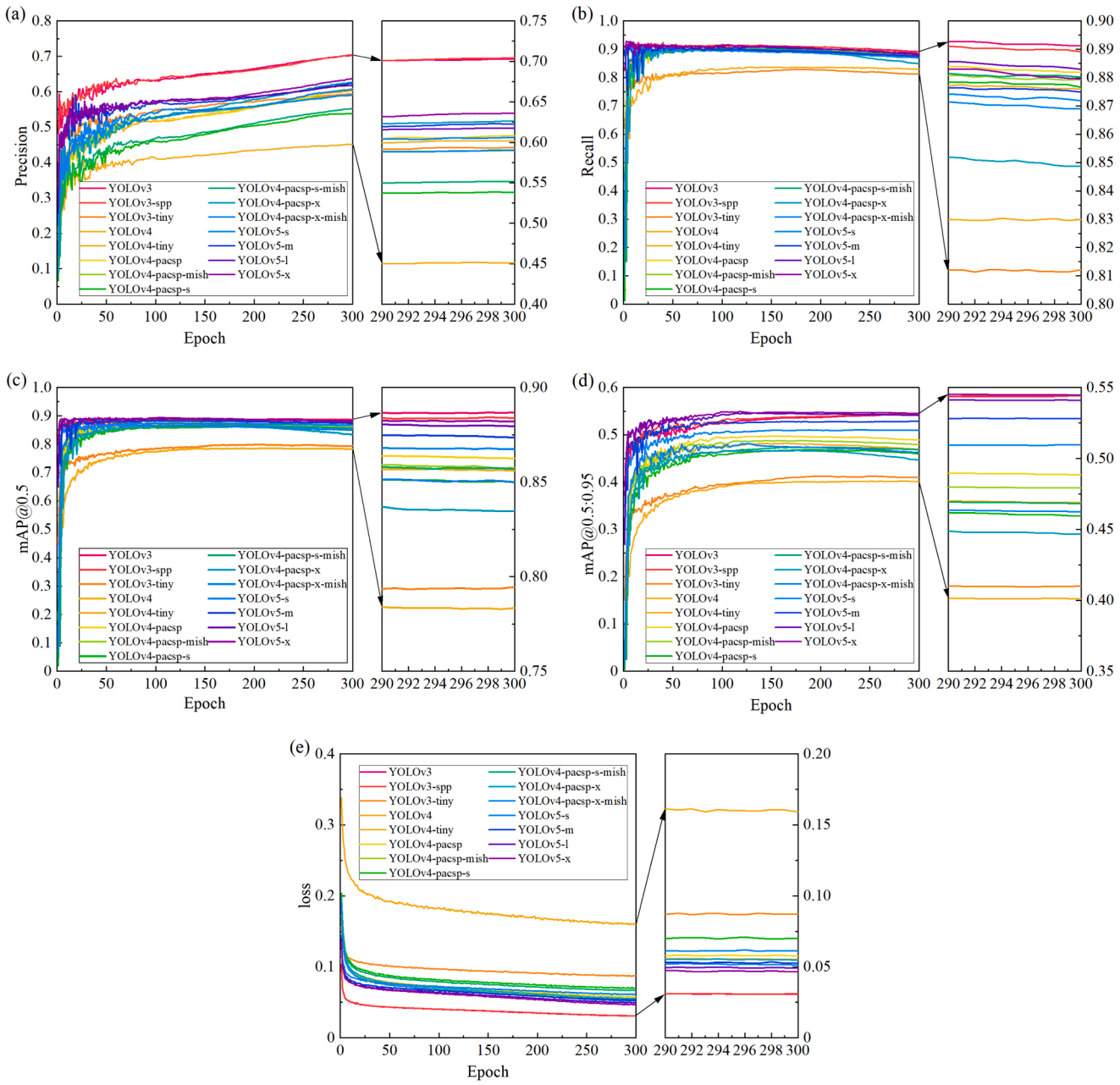


Fig. 11. Retraining results of YOLOv3, YOLOv4 and YOLOv5 with different architectures. (a) Training precision of different models; (b) Training recall of different models; (c) The mAP@0.5 scores of different models; (d) The mAP@0.5:0.95 scores of different models; (e) The loss curves of the training dataset.

results. For the YOLOv4 model, YOLOv4-pacsp and YOLOv4-pacsp-mish had the highest recall scores at 0.697. The overall precision of the YOLOv4 models was lower than that of the other algorithms because the YOLOv4 models had false positive detections that contributed to the low precision of the testing results. Additionally, YOLOv5-s had the highest precision and mAP for cars and persons in comparison with the other architectures in YOLOv5 models. In terms of speed, smaller and simpler network detection is faster than larger and more complicated network detection. YOLOv3-tiny, YOLOv4-tiny, and YOLOv5-s had a faster detection speed compared with the other architectures. Among them, YOLOv3-tiny had the fastest detection speed in the 833 video TIR testing images in the testing FLIR dataset at 50 FPS. According to the best model storage evaluation results, the YOLOv5 models' storage (i.e., 14 MB) did not exceed 200 MB and occupied the least storage in the computer among all algorithms. YOLOv4-pacsp-x and YOLOv4-pacsp-x-mish had the largest storage at 379 MB. The YOLOv3 algorithms with high detection accuracy had a storage capacity of about 100 MB in the best-retrained model.

For the qualitative assessment of detection image results, we compared the models with better performance of the COCO testing set than the others (Ultralytics, 2020a, 2020b; Wong, 2020). Fig. 12 illustrates the detection results of FLIR training images from four scenarios based on YOLOv3-spp, YOLOv4-pacsp-x-mish and YOLOv5-x, respectively.

The performance in different methods and network architecture conditions showed that some missed detection incidents occurred in the YOLOv3 models. These models did not respond satisfactorily to the person instances that appear small in the testing image of Fig. 12 (b1), and the small cars in the testing image of Fig. 12 (e1), which is of myriad importance when the models are used in public security fields, such as finding missing people or even criminals in mission-critical case scenarios and detecting pedestrians in ADAS. Additionally, in the clustered

Table 2
Multi-object detection results of YOLOv3, YOLOv4 and YOLOv5 with different architectures in FLIR testing video images. The best and second-best scores are **highlighted** and underlined.

	Models	Average Precision of Car	Average Precision of Person	mAP	Precision	Recall	Speed (FPS)	Storage (MB)
YOLOv3	YOLOv3	0.904	**0.739**	0.821	0.712	0.677	30	117
	YOLOv3-spp	0.909	0.734	**0.822**	**0.718**	0.671	30	119
	YOLOv3-tiny	0.893	0.600	0.747	0.613	0.603	**50**	17
YOLOv4	YOLOv4	0.846	0.371	0.608	0.600	0.690	24	244
	YOLOv4-tiny	0.794	0.241	0.517	0.525	0.595	45	22
	YOLOv4-pacsp	0.848	0.392	0.620	0.607	**0.697**	21	200
	YOLOv4-pacsp-mish	0.853	0.390	0.622	0.604	**0.697**	27	200
	YOLOv4-pacsp-s	0.849	0.360	0.604	0.591	0.680	32	31
	YOLOv4-pacsp-s-mish	0.849	0.363	0.606	0.593	0.684	33	31
	YOLOv4-pacsp-x	0.833	0.337	0.585	0.596	0.667	23	379
	YOLOv4-pacsp-x-mish	0.848	0.358	0.603	0.606	0.681	22	379
YOLOv5	YOLOv5-s	**0.918**	0.688	0.803	0.638	0.670	41	**14**
	YOLOv5-m	0.915	0.677	0.796	0.601	0.692	35	41
	YOLOv5-l	0.912	0.674	0.793	0.619	0.690	28	91
	YOLOv5-x	0.913	0.671	0.792	0.621	0.687	23	169

stopping cars (Fig. 12 (a0)), 13 predicted bounding boxes that have equal numbers of annotations in Fig. 12 (a1) were detected by the YOLOv3-spp method, but the accuracy of small car instance detection was not too high. Evidently, the YOLOv3-spp model shows better performance on large-scale objects, such as the cars in the FLIR testing set, compared with person instances.

In accordance with the YOLOv4-pacsp-x-mish results, false positive detections can be observed in the 3rd column in Fig. 12 related to the architecture of the YOLOv4 model, especially for the car instances. The YOLOv5-x detection results revealed that a similar false positive detection in the multi-objects case was obtained on either side of the street scenario. YOLOv3 outperformed the other models, but a missed detection by YOLOv3-spp was observed in the pedestrian scenario (the smallest person was not detected), which is called false negative detection.

Based on the training models, the testing results of four models, including YOLOv4-pacsp-x-mish, YOLOv4-pacsp, YOLOv5-x, and YOLOv5-s, are presented in Fig. 13. YOLOv4-pacsp had a false positive detection for the ride person in Fig. 13 (b1). The cyclist detection results obtained with YOLOv4-pacsp-x-mish were similar to those obtained with YOLOv5-x and YOLOv5-s, which can detect the cyclist successfully and accurately. The comparison of Fig. 13 (b3) and Fig. 13 (b4) for the person instances that appear small in the image revealed that YOLOv5-x could detect more small objects than YOLOv5-s, but it showed a disadvantage in the false positive detections in Fig. 13 (d3). A number of non-person instances were identified as persons.

The testing experiment revealed that in the clustered stopping car images, such as in Fig. 12 (a0), most of the models had false-positive detections, except for YOLOv3-spp. With regards to the cyclist detection in Fig. 13 (b0), only YOLOv4-pacsp had a false positive detection, namely, the rider was detected as two persons. In the moving car scene with no overlapping areas at intersections in Fig. 12 (c0) and Fig. 13 (c0), all methods perfectly detected the moving cars. In the scenes where only a person instance or persons and cars were present on either side of the street, such as in Fig. 12 (d0) (or Fig. 13 (d0)) and Fig. 12 (e0) (or Fig. 13 (e0)), YOLOv4 and YOLOv5 had false positive detections. Meanwhile, person instances that appear small in Fig. 12 were missed detected by the YOLOv3-spp model. Overall, these results confirm the successful and highly sophisticated real-time detection capability of YOLO methods in multi-object and multi-scenario cases.

In conclusion, considering the optimal balance of the quantitative assessment with evaluation metrics, we chose the YOLOv5-s model to realize the detection of UAV TIR videos. YOLOv5-s model has the smallest size and relatively precision and speed of detection. A slim model would be more practical in video detection with mobile devices. Moreover, we also discussed the detection effect of the YOLOv3-spp model which has the highest mAP on validation tasks and the YOLOv5-s model on UAV thermal infrared video in the section of the discussion.

5. Cross-application in UAV TIR videos

During the testing stage, unlabeled UAV TIR videos were detected based on the training model in the ground TIR video images from different sensors. With different zenith angles, UAV TIR videos with 6402 frames of oblique images were fed in the YOLOv5-s retrained model. The key parameters of testing detection were empirically set as confidence threshold = 0.25 and IoU = 0.65 to correspond with the detection task in FLIR TIR video images. The output produced the location, confidence, and class category in the video. We analyzed the accuracy of the UAV TIR video testing results from qualitative and quantitative perspectives, including video results and detection speed.

From the detection results in Fig. 14 (a) (the corresponding result video is provided as a supplementary document), we can infer that the persons in the low-height thermal aerial videos can be successfully detected using YOLO model, which is trained from ground TIR images. The number of persons detected fluctuated within the blue lines from 4 to 8, which is consistent with the actual number. Therefore, in the low flight height UAV TIR video with crowded persons, the flight movement of the UAV has little effect on the fluctuation of the walking person detection result in this study.

However, in the UAV TIR video with high flight height, the detection results of stopping cars and moving persons fluctuated greatly with the UAV forward (Fig. 14 (b)). All cars were detected well in most frames, but only one car was detected in the frames under large flight changes. By contrast, moving persons occupied a relatively small portion of the entire image and included large complex environment information, which increased the detection complexity. Thus, if the features of the person were not obvious, it would be difficult to detect a moving person.

FPS, as inference detection time, was also used to evaluate the speed performance. UAV TIR videos with 6233 and 169 frames were detected in 154.495 s and 4.195 s, respectively. As shown in Fig. 15, we used one millisecond as the step size to count the number of frames in each interval separately and found that the hot spots were concentrated between 0.0105 and 0.0125 s (Fig. 15 (a)). As seen in Fig. 15 (b), the detection speed heat map is sparsely distributed because the total number of frames was small. The detection speed was relatively clustered between 0.010 and 0.013 s. These results confirm that UAV TIR videos can achieve real-time performance with 40 FPS.

Overall, this application on UAV TIR videos proved the scalability and flexibility of the model trained on ground TIR images. Taking advantage of the remarkable real-time detection capability of the YOLO

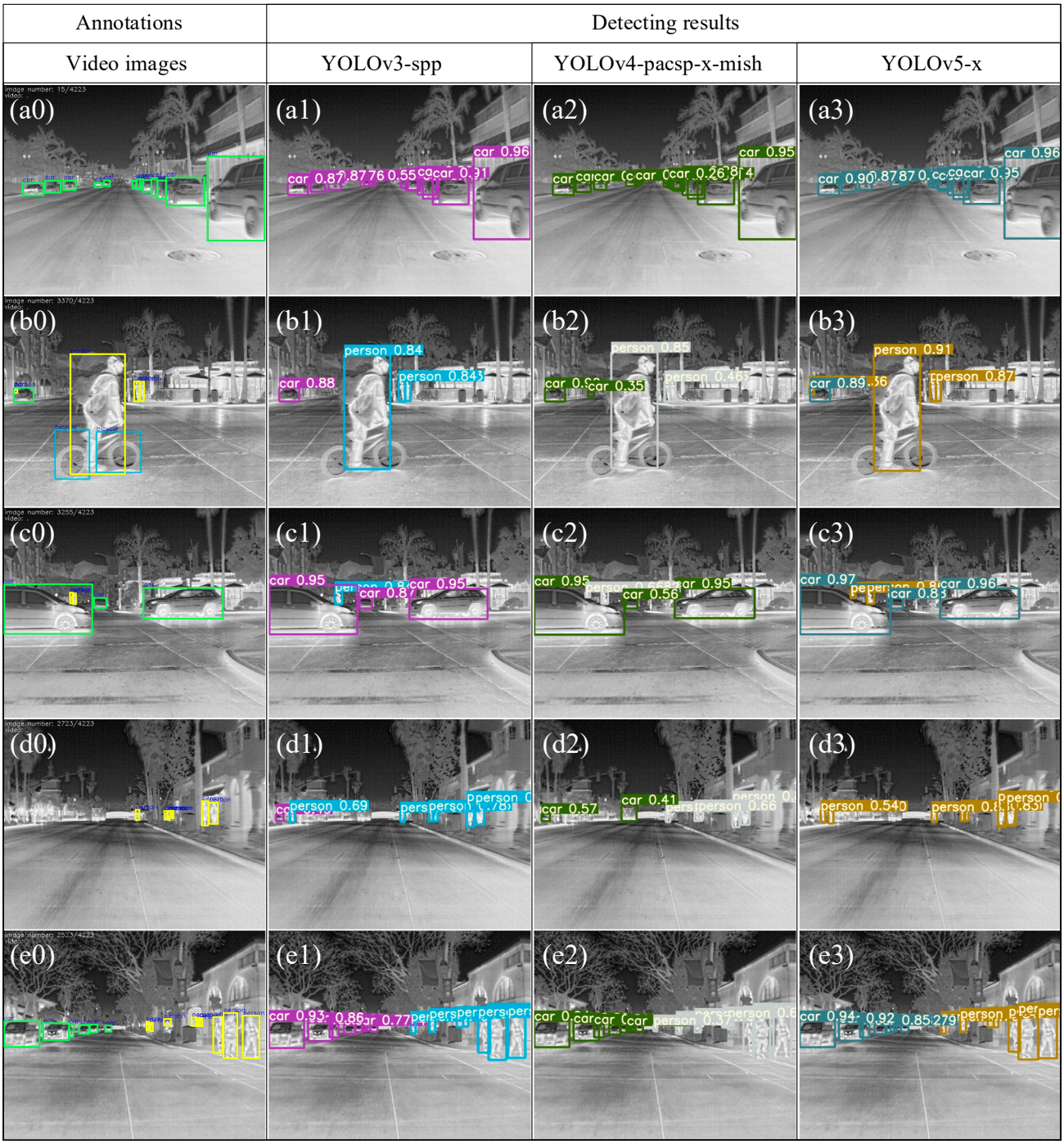


Fig. 12. Illustrations of detection results on FLIR video images. (a0)-(e0)First row images are the original label image. The remaining three rows of images are the detection results of the YOLOv3-spp, YOLOv4-pacsp-x-mish and YOLOv5-x models, respectively.

model in the complex, changeable background on UAV TIR videos, this model can assist in real-time emergency monitoring and offers the most efficient and operative means to achieve full-time, high-precision UAV patrol monitoring.

6. Discussion and conclusion

6.1. Discussion

Due to the restrictions of remote control aircraft and three-axis motion, object detection from UAV images and videos has always been an important but challenging task in order to improve the autonomy of UAV and remote sensing applications (Aposporis, 2020). This study performed multi-object detection experiments using the YOLO models with 15 different architectures based on pre-trained checkpoints from multi-scenario ground TIR video images. In order to search for a relative outstanding performance model, this study validated the training models on the ground TIR dataset with evaluation metrics. The best model was used for various UAV-based TIR video detection under a natural environment, thus realizing transfer detection from ground TIR video images to UAV TIR videos.

In the training and validation task, pre-processed FLIR datasets with multiple objects and scenarios were fed into the YOLOv3, YOLOv4, and YOLOv5 architectures with 15 models for speed and accuracy

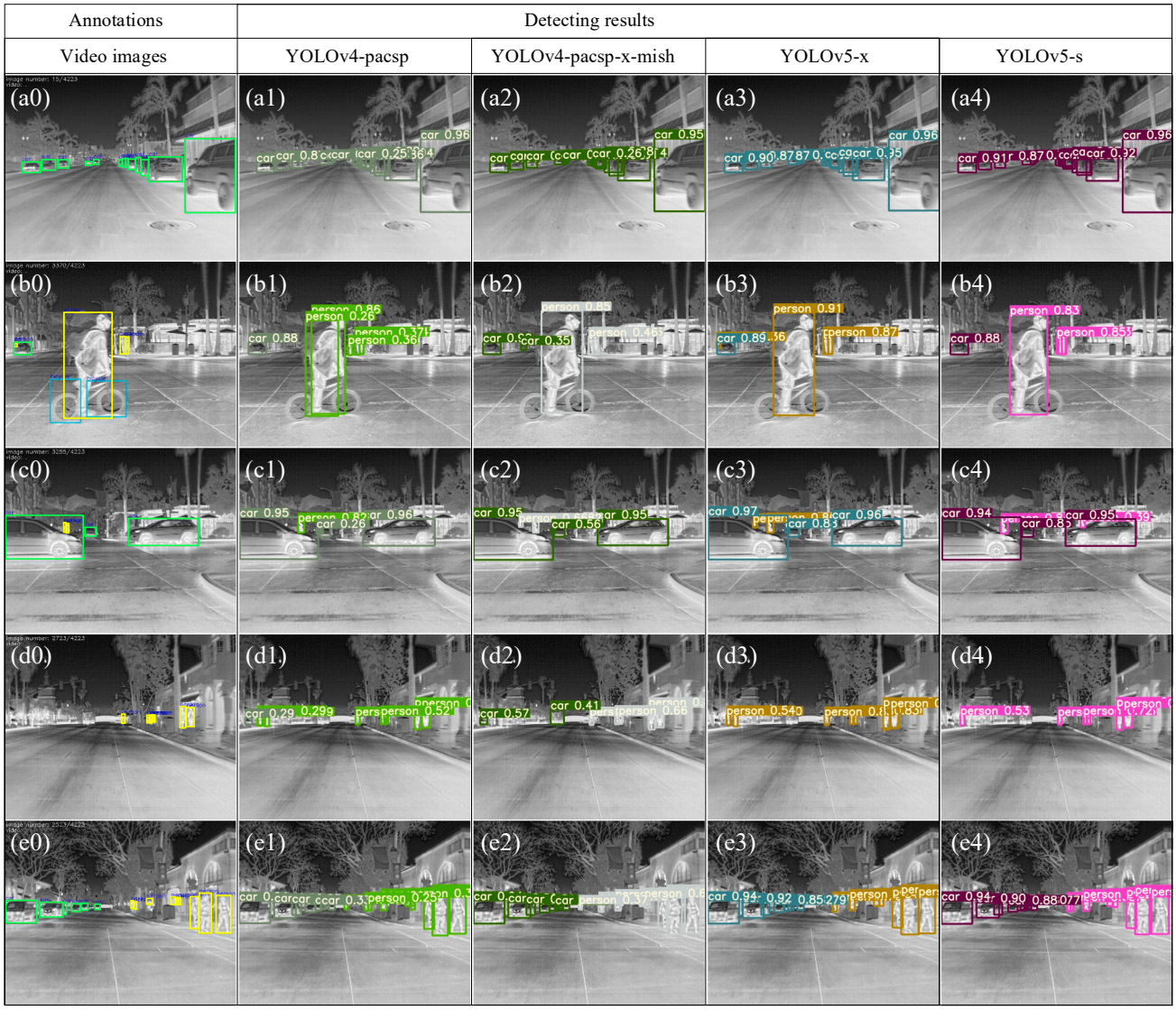


Fig. 13. Comparisons results of YOLOv4-pacsp and YOLOv4-pacsp-x-mish, YOLOv5-s and YOLOv5-x.

evaluation. Training experiments showed that the overall recall score was above 80 %, and the precision score was between 45 % and 80 %. The precision scores of YOLOv3-spp and standard YOLOv3 (70.38 % and 70.31 %, respectively) were superior to those of the other models. The comparison of training validation data by using the mAP evaluation index showed that the YOLOv3 achieved the highest score of mAP@0.5 at 88.69 % in the validation task of detecting cars and persons. The mAP@0.5 scores of the YOLOv5 models were extremely close to 87 %, and the scores of mAP@0.5 in the YOLOv4 architectures (except simple architecture) ranged between 83 % and 87 %. Additionally, we conducted a comparison of mAP@0.5:0.95 with 0.05 step size on the training validation data and found that the most sophisticated stochastic approach in the score of mAP@0.5:0.95 was YOLOv5-x with 54.46 %. The simple architectures used in YOLOv3 and YOLOv4 had a lower precision of mAP@0.5:0.95 compared with the other models.

Moreover, in the testing task, the retrained models were tested to detect objects in TIR video images through evaluation metrics. The YOLOv3 architectures achieved the generally highest precision and mAP scores for all classes at 0.718 and 0.822, respectively. In each class detection, YOLOv3, YOLOv4-pacsp-mish, and YOLOv5-s showed better performance than the other tiny models in their network architectures with regard to person instance detection. The small and simple architectures, including YOLOv3-tiny, YOLOv4-tiny, and YOLOv5-s, achieved faster detection. The highest detection speed of 50 FPS was observed for YOLOv3-tiny for the 833 TIR video images, and YOLOv3-tiny had good detection results for a large instance. In terms of the best model storage, YOLOv4-pacsp-x and YOLOv4-pacsp-x-mish were slow at 23 and 22 FPS, respectively, and had more storage size at 379 MB. However, the YOLOv5 models' storage size did not exceed 200 MB, and the least storage was for YOLOv5-s with 14 MB. In consideration of a trade-off between the precisions, the YOLOv3 and YOLOv5 showed better performance among all the YOLO approaches for both car and person detection.

In consideration of the speed and storage metrics, the YOLOv5 models (especially YOLOv5-s) had better performance on detecting TIR video images from in-vehicle cameras. Considering the speed of real-time detection, the random access memory (RAM) in the UAV platform, and the overall accuracy, we tested YOLOv5-s suitable for the long sequences of UAV TIR videos. We also tested the YOLOv3-spp and YOLOv5-s in a 4-min UAV TIR video. YOLOv3-spp showed the best mAP performance in the validation task, but it has false positive samples in UAV TIR video detection. As shown in Fig. 16 (a) and (b), the model identifies the UAV equipment on the ground as a person, which will lead to the overhead of additional computing power and storage space. By contrast, with the same confidence, YOLOv5-s could better detect persons, as shown in Fig. 16 (b) and (c) below. Moreover, the detection

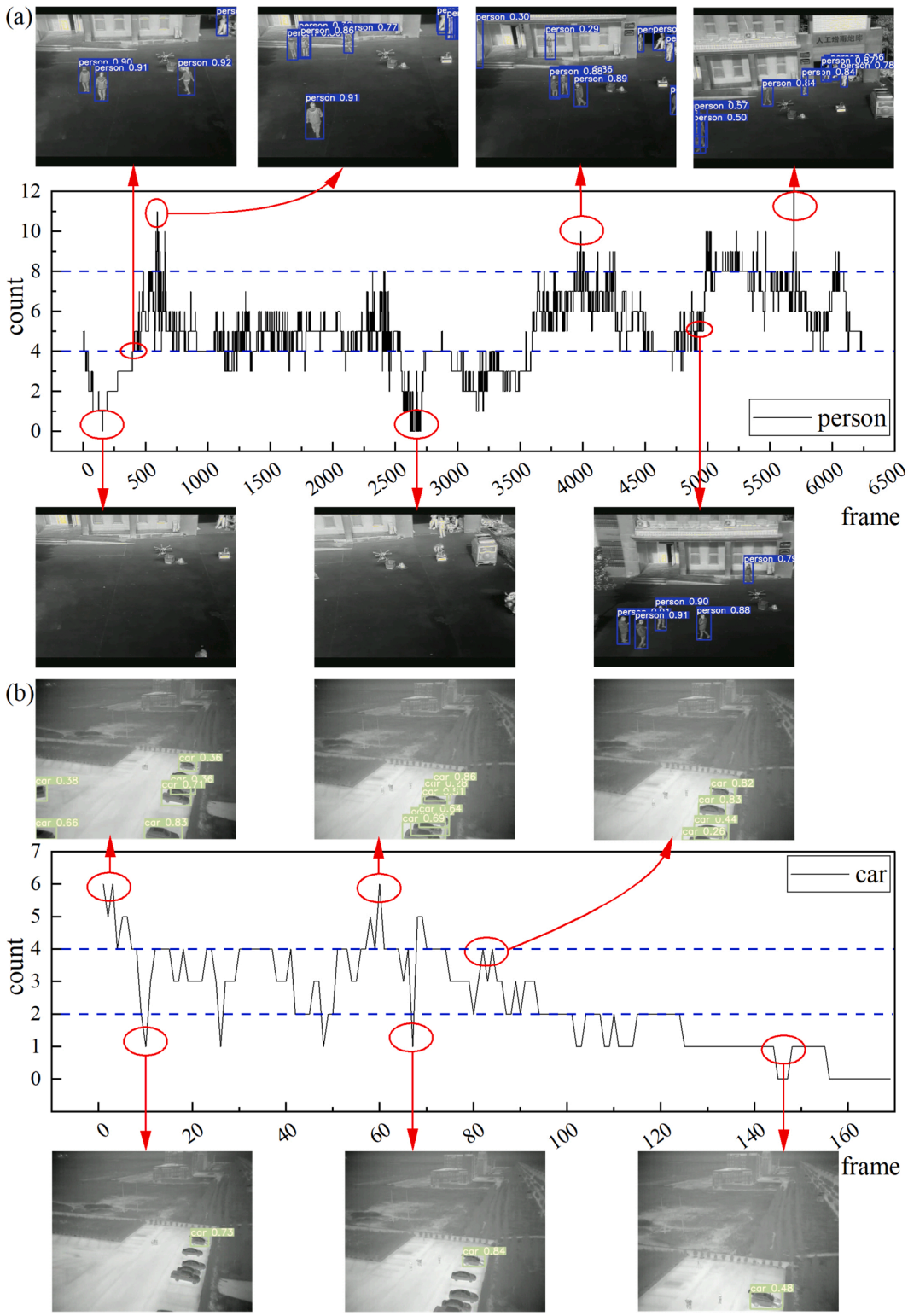


Fig. 14. Evaluation of UAV TIR videos detection results. (a) TIR video frame of UAV at the lower flight height; (b) TIR video frame of UAV at the higher flight height.

speed of the YOLOv3-spp and YOLOv5-s are 111.819 s and 105.921 s, respectively, in the 4-min UAV TIR video.

In cross-application, two customized UAV TIR videos with different zenith angles were adopted in this study. At low flight height, the UAV TIR video object detection performed well, but as the UAV flight angle and height increased, the detection accuracy fluctuated greatly. This result may be attributed to the similar information between ground videos and UAV videos. At low flight height, high similarity information was retained. However, as the flying height increased, the information on persons and cars changed significantly. UAV TIR video detection could achieve real-time performance with 40 FPS. The detection results proved that YOLO models could be successfully applied in low flight height UAV TIR video for object detection and are not limited to detecting images captured from in-vehicle cameras.

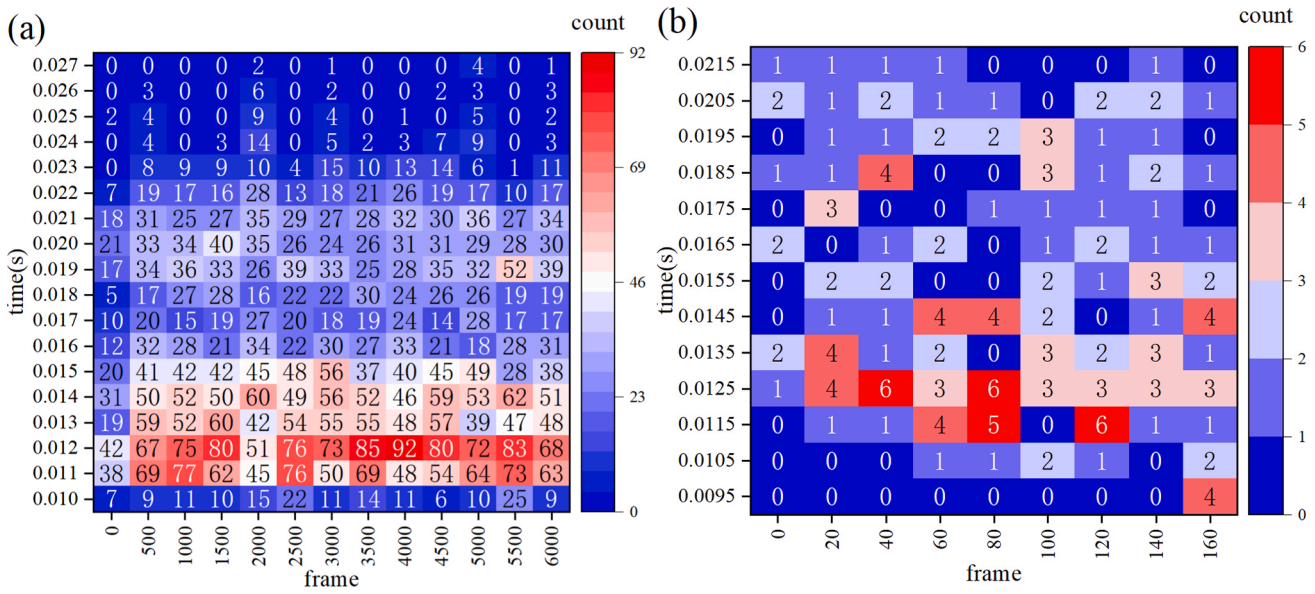


Fig.15. UAVs video real-time monitoring heat map.

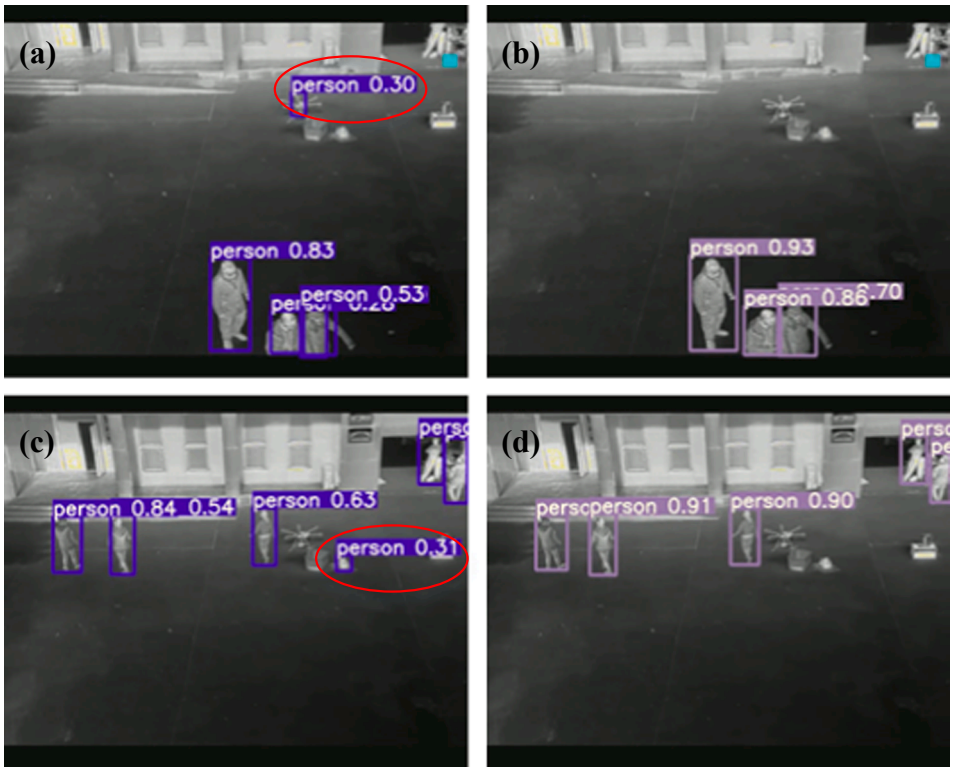


Fig. 16. Detection results of YOLOv3-spp and YOLOv5-s from the UAV TIR video. (a) and (c) YOLO-spp detection results; (b) and (d) YOLOv5-s detection results.

6.2. Conclusion

The UAV object detection framework based on YOLO models in this research was proven to be efficient in detecting TIR images and videos with multiple objects. This study performed car and person multi-object detection experiments using the YOLO models (i.e., YOLOv3, YOLOv4, and YOLOv5) with 15 different architectures based on pre-trained checkpoints from multi-scenario ground TIR video images. The highest mAP of all classes was 88.69 %, the fastest detection speed achieved 50 FPS, and the smallest model size observed in YOLOv5-s was 14 MB in the validating task. The UAV object-based detection framework we proposed has taken advantage of the YOLO models trained based on the ground TIR dataset. In order to get satisfactory detection results, we also discussed the selection of evaluation metrics. The framework we proposed to realize the object detection was rather a routine than some separate steps. It is expected that object detection in other TIR datasets can benefit from the detection results and cross-application as presented in this study.

Further research can contribute to the additional use of new UAV TIR video datasets and the development of novel deep learning architectures for UAV TIR object detection. The optimum viewing angle for dealing with the effect of the observation angle in UAV is also required

investigation in order to detect quickly and exactly.

Author statement

Declaration of competing interest and funding: we confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Author agreement statement: we confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

Protect intellectual property: we confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 42130104), the National High-Resolution Earth Observation Project (No. 30-H30C01-9004-19/21), the National Natural Science Foundation of China (No. 41771369 and 42101340), China Postdoctoral Science Foundation under Grant (No. 2021M690199).

References

Aposporis, P., 2020. Object detection methods for improving UAV autonomy and remote sensing applications. Proc. IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 845–853 https://doi.org/10.1109/ASONAM49781.2020.9381377.

Bochkovskiy, A., Wang, C., Liao, H.M., 2020. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934. doi:10.48550/arXiv.2004.10934.

Chen, Y., Zhang, C., Qiao, T., Xiong, J., Liu, B., 2021. Ship detection in optical sensing images based on YOLOv5. Proc. SPIE Int. Conf. Graph. Image Process. 11720, 117200E–1–117200E–5. doi:10.1117/12.2589395.

Chen, Y., Shin, H., 2020. Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network. Appl. Sci. 10, 1–18. https://doi.org/10.3390/app10030809.

Farouk Khalifa, A., Badr, E., Elmahdy, H.N., 2019. A survey on human detection surveillance systems for Raspberry Pi. Image Vis. Comput. 85, 1–13. https://doi.org/10.1016/j.imavis.2019.02.010.

Flir, 2019. FLIR thermal starter dataset introduction version 1.3. /adasdataset (accessed Aug. 16, 2019). https://www.flir.com.

Garrido, I., Lagüela, S., Arias, P., Balado, J., 2018. Thermal-based analysis for the automatic detection and characterization of thermal bridges in buildings. Energy Build. 158, 1358–1367. https://doi.org/10.1016/j.enbuild.2017.11.031.

Girshick, R., 2015. Fast R-CNN. Proc. IEEE Int. Conf. Comput. Vis. 1440–1448 https://doi.org/10.1109/ICCV.2015.169.

Girshick, R., Donahue, J., Darrell, T., Malik, J., Berkeley, U.C., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 580–587. doi:10.1109/CVPR.2014.81.

Gong, X., Yao, Q., Wang, M., Lin, Y., 2018. A deep learning approach for oriented electrical equipment detection in thermal images. IEEE Access 6, 41590–41597. https://doi.org/10.1109/ACCESS.2018.2859048.

Hassan, S.A., Rahim, T., Shin, S.Y., 2019. Real-time UAV detection based on deep learning network. Proc. Int. Conf. Inf. Commun. Technol. Converg. 630–632 https://doi.org/10.1109/ICTC46691.2019.8939564.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824.

Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., Wang, R., 2020. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Inf. Sci. 522, 241–258. https://doi.org/10.1016/j.ins.2020.02.067.

Iwasaki, Y., Misumi, M., Nakamiya, T., 2013. Robust vehicle detection under various environmental conditions using an infrared thermal camera and its application to road traffic flow monitoring. Sensors 13, 7756–7773. https://doi.org/10.3390/s130607756.

Kanistras, K., Martins, G., Rutherford, M.J., Valavanis, K.P., 2013. A survey of unmanned aerial vehicles (UAVs) for traffic monitoring. Proc. Int. Conf. Unmanned Aircr. Syst. 221–234 https://doi.org/10.1109/ICUAS.2013.6564694.

Kannadaguli, P., 2020. YOLOv4 based human detection system using aerial thermal imaging for UAV based surveillance applications. Proc. IEEE Int. Conf. Decis. Aid Sci. Appl. 1213–1219. doi:10.1109/DASA51403.2020.9317198.

Khanal, S., Fulton, J., Shearer, S., 2017. An overview of current and potential applications of thermal remote sensing in precision agriculture. Comput. Electron. Agric. 139, 22–32. https://doi.org/10.1016/j.compag.2017.05.001.

Khoshboresh Masouleh, M., Shah-Hosseini, R., 2019. Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from UAV-based thermal infrared imagery. ISPRS J. Photogramm. Remote Sens. 155, 172–186. https://doi.org/10.1016/j.isprsjprs.2019.07.009.

Kristo, M., Ivasic-Kos, M., Pobar, M., 2020. Thermal Object Detection in Difficult Weather Conditions Using YOLO. IEEE Access 8, 125459–125476.

Kumar, S., Yadav, D., Gupta, H., Verma, O.P., Ansari, I.A., Ahn, C.W., 2021. A novel YOLOv3 algorithm-based deep learning approach for waste segregation: Towards smart waste management. Electron. 10, 1–20. https://doi.org/10.3390/electronics10010014.

Lee, J., Wang, J., Crandall, D., Šabanović, S., Fox, G., 2017. Real-Time, cloud-based object detection for unmanned aerial vehicles. Proc. IEEE Int. Conf. Robot. Comput. 36–43 https://doi.org/10.1109/IRC.2017.77.

Leira, F.S., Johansen, T.A., Fossen, T.I., 2015. Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera. Proc. IEEE Aerosp. Conf. 1–10 https://doi.org/10.1109/AERO.2015.7119238.

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. 159, 296–307. https://doi.org/10.1016/j.isprsjprs.2019.11.023.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. Proc. IEEE Conf. Comput. Vis. pattern Recognit. 936–944 https://doi.org/10.1109/CVPR.2017.106.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. Proc. IEEE Int. Conf. Comput. Vis. 2980–2988. doi:10.1109/ICCV.2017.324.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. Proc. Eur. Conf. Comput. Vis. 740–755. https://doi.org/10.48550/arXiv.1405.0312.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. SSD: Single shot multibox detector. Proc. Eur. Conf. Comput. Vis. 21–37 https://doi.org/10.1007/978-3-319-46448-0_2.

Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., Piao, C., 2020. UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. Sensors 20, 1–12. doi:10.3390/s20082238.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts. Proc. 5th Int. Conf. Learn. Represent. 1–16. doi:10.48550/ arXiv.1608.03983.

Misra, D., 2019. Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv: 1908.08681. doi:10.48550/arXiv.1908.08681.

Redmon, J., Farhadi, A., 2018. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767. doi:10.48550/arXiv.1804.02767.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. 779–788 https://doi.org/10.1021/je00029a022.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

Ren, X., Sun, M., Zhang, X., Liu, L., Zhou, H., Ren, X., 2022. An improved mask-RCNN algorithm for UAV TIR video stream target detection. Int. J. Appl. Earth Obs. Geoinf. 106, 102660. https://doi.org/10.1016/j.jag.2021.102660.

Rudol, P., Doherty, P., 2008. Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery. IEEE Aerosp. Conf. 1–8 https://doi.org/10.1109/AERO.2008.4526559.

Shao, Z., Cheng, G., Ma, J., Wang, Z., Wang, J., Li, D., 2021. Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. IEEE Trans. Multimed. 1–16 https://doi.org/10.1109/TMM.2021. 3075566.

Shin, D.-J., Kim, J.-J., 2022. A deep learning framework performance evaluation to use YOLO in nvidia jetson platform. Appl. Sci. 12 (8), 3734. https://doi.org/10.3390/app12083734.

Tian, D., Lin, C., Zhou, J., Duan, X., Cao, Y., Zhao, D., Cao, D., 2020. SA-YOLOv3: An efficient and accurate object detector using self-attention mechanism for autonomous driving. IEEE Trans. Intell. Transp. Syst. 1–12. doi:10.1109/TITS.2020.3041278.

Ultralytics, 2020a. YOLOv5. https://github.com/ultralytics/yolov5 (accessed Oct. 29, 2020).

Ultralytics, 2020b. YOLOv3. https://github.com/ultralytics/yolov3 (accessed Dec. 7, 2020).

Wang, C.Y., Mark Liao, H.Y., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H., 2020. CSPNet: A new backbone that can enhance learning capability of CNN. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 1571–1580. doi:10.1109/CVPRW50498.2020.00203.

Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J., 2019. PANet: Few-shot image semantic segmentation with prototype alignment. Proc. IEEE/CVF Int. Conf. Comput. Vis. 9196–9205 https://doi.org/10.1109/ICCV.2019.00929.

Wong, K.-Y., 2020. PyTorch_YOLOv4. https://github.com/WongKinYiu/PyTorch_YOLOv4 (accessed Nov. 26, 2020).

Wu, D., Lv, S., Jiang, M., Song, H., 2020. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. Computers and Electronics in Agriculture 178, 105742. https://doi.org/10.1016/j.compag.2020.105742.

Xiang, T.-Z., Xia, G.-S., Zhang, L., 2019. Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects. IEEE Geosci. Remote Sens. Mag. 7 (3), 29–63. https://doi.org/10.1109/MGRS.2019.2918840.

Xu, Z., Zhuang, J., Liu, Q., Zhou, J., Peng, S., 2019. Benchmarking a large-scale FIR dataset for on-road pedestrian detection. Infrared Phys. Technol. 96, 199–208. https://doi.org/10.1016/j.infrared.2018.11.007.

Zhang, H., Luo, C., Wang, Q., Kitchin, M., Parmley, A., Monge-Alvarez, J., Casaseca-de-la-Higuera, P., 2018. A novel infrared video surveillance system using deep learning based techniques. Multimed. Tools Appl. 77, 26657–26676. https://doi.org/10.1007/s11042-018-5883-y.

Zhang, P., Zhong, Y., Li, X., 2019. SlimYOLOv3: Narrower, faster and better for real-time UAV applications. Proc. IEEE/CVF Int. Conf. Comput. Vis. 1–11 https://doi.org/10.1109/ICCVW.2019.00011.

Zhao, Q., Sheng, T., Wang, Y., Ni, F., Cai, L., 2018. CFENet: An accurate and efficient single-shot object detector for autonomous driving. arXiv preprint arXiv: 1806.09790. doi:10.48550/arXiv.1806.09790.

Zhao, L., Li, S., 2020. Object detection algorithm based on improved YOLOv3. Electron. 9, 1–11. https://doi.org/10.3390/electronics9030537.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-IoU loss: Faster and better learning for bounding box regression. Proc. AAAI Conf. Artif. Intell. 12993–13000. https://doi.org/10.48550/arXiv.1911.08287.

Zhong, Y., Hu, X., Luo, C., Wang, X., Zhao, J., Zhang, L., 2020. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. Remote Sens. Environ. 250, 112012 https://doi.org/10.1016/j.rse. 2020.112012.

Zou, Z., Shi, Z., Guo, Y., Ye, J., 2019. Object detection in 20 years: A survey. arXiv preprint arXiv: 1905.05055. doi:10.48550/arXiv.1905.05055.