remote sensing

MDPI

Article

A Survey of Object Detection for UAVs Based on Deep Learning

Guangyi Tang [1], Jianjun Ni [1,2,*], Yonghao Zhao [1], Yang Gu [1] and Weidong Cao [1,2]

[1] College of Artificial Intelligence and Automation, Hohai University, Changzhou 213200, China; tang_gy@hhu.edu.cn (G.T.); zhaoyoh@hhu.edu.cn (Y.Z.); guyang.passdom@gmail.com (Y.G.); cwd2018@hhu.edu.cn (W.C.)

[2] College of Information Science and Engineering, Hohai University, Changzhou 213200, China

* Correspondence: njjhhuc@gmail.com

Abstract: With the rapid development of object detection technology for unmanned aerial vehicles (UAVs), it is convenient to collect data from UAV aerial photographs. They have a wide range of applications in several fields, such as monitoring, geological exploration, precision agriculture, and disaster early warning. In recent years, many methods based on artificial intelligence have been proposed for UAV object detection, and deep learning is a key area in this field. Significant progress has been achieved in the area of deep-learning-based UAV object detection. Thus, this paper presents a review of recent research on deep-learning-based UAV object detection. This survey provides an overview of the development of UAVs and summarizes the deep-learning-based methods in object detection for UAVs. In addition, the key issues in UAV object detection are analyzed, such as small object detection, object detection under complex backgrounds, object rotation, scale change, and category imbalance problems. Then, some representative solutions based on deep learning for these issues are summarized. Finally, future research directions in the field of UAV object detection are discussed.

Keywords: object detection; unmanned aerial vehicles; deep learning; computer vision

Citation: Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149. https://doi.org/10.3390/rs16010149

Academic Editor: Pedro Melo-Pinto

Received: 22 November 2023
Revised: 26 December 2023
Accepted: 27 December 2023
Published: 29 December 2023

1. Introduction

Object detection is an important research topic in the field of remote sensing, which has been widely applied to military and civil tasks, such as geological environment surveys, traffic monitoring, urban planning, precision agriculture, and disaster relief [1–4]. Traditional methods of obtaining objects mainly rely on satellites and manned aircraft. In recent years, since unmanned aerial vehicles (UAVs) have the advantages of being small, flexible, and easy to control, they have become more and more popular in various domains, such as civilian, military, and research. Some examples are inspecting power lines in difficult conditions, detecting air quality, rescuing people in danger, spying on the enemy, tracking enemy targets, and searching for information on the battlefield. All these tasks strongly depend on the detection of one or more domain-specific objects. Since the emergence of computer vision, object detection has been extensively researched.

Historically, detecting objects in images captured by cameras is one of the earliest computer vision tasks, dating back to the 1960s. Detecting and recognizing object categories has been viewed as the key element of any artificial intelligence system ever since. Many computer vision methods have been used for this task.

As one of the basic problems for computer vision, traditional object detection adopted statistical-based methods [5]. However, the massive amount of data has affected the performance of these traditional methods in recent years. Problems such as feature size explosion, requiring larger storage space and time costs, are difficult to solve. With the emergence of deep neural network (deep learning) technology [6–8], high-level features of images can be extracted through multiple convolutions and pooling layers, thus achieving object detection. Since 2015, deep neural networks have become the main framework for

UAV object detection [9,10]. Classical deep neural networks for object detection are divided into two categories: two-stage networks and one-stage networks. Two-stage networks, such as RCNN [11] and Faster RCNN [12,13], first need to generate proposal regions and then classify and locate the proposal regions. Numerous studies have demonstrated that two-stage networks are appropriate for applications with higher detection accuracy requirements [14,15]. One-stage networks, such as SSD [16,17] and Yolo [18,19], directly generate class probabilities and coordinate positions and are faster than two-stage networks. Therefore, one-stage networks have great advantages in UAV practical applications with high-speed requirements. Similarly, there are also some faster lightweight networks, such as MobileNet SSD [20], Yolov3 [21], ESPNetv2 [22], etc.

Object detection for UAVs based on deep learning can acquire and analyze information about the ground scene in real time during flight, thus improving the perception and intelligence of UAVs. However, there are many new challenges regarding UAV object detection compared to ground object detection, such as the low-quality image problem and the complex background problem. Therefore, this paper provides a survey on object detection for UAVs based on deep learning. Firstly, we analyze the types and sensors of UAVs. Then, we point out the main differences between UAV object detection and common object detection, and the main challenges in UAV object detection. On this basis, we focus on the applications of deep learning for UAV object detection from the view of the challenges. Other relevant surveys in the field of object detection based on deep learning can be used as a supplement to this paper (see, e.g., [23–26]).

The main contributions of this paper are summarized as follows: (1) the development of UAV object detection is thoroughly analyzed and reviewed. In addition, the differences between UAV object detection and common object detection are analyzed and the challenges in UAV object detection are enumerated. (2) A survey on the applications of deep learning methods in UAV object detection is provided, which focuses on the main challenges in UAV object detection. (3) Some representative methods in the field of UAV object detection based on deep learning are analyzed. At last, some possible future study directions in this field are discussed.

The rest of this paper is organized as follows: Section 2 briefly outlines the development history of UAV object detection. Section 3 describes the deep-learning-based object detection algorithms for UAV aerial images. Section 4 surveys the main public datasets used for UAV object detection. Section 5 combines the current research status to look forward to the follow-up research direction, and Section 6 is a conclusion.

2. The Development of UAV Object Detection

The increase in remote sensing systems enables people to collect data regarding any objects on Earth's surface extensively. With the emergence of UAVs, aerial imaging has become a common method of data acquisition.

2.1. Classification of UAVs

Different criteria can be used to categorize UAVs, such as lift system, payload weight, size, maximum takeoff weight, operational range, operational altitude (above ground level), endurance, operational conditions, or autonomy level [27]. Based on the lift system, they can be divided into several types, such as lighter-than-air UAVs, fixed-wing UAVs, rotary-wing UAVs, flapping-wing UAVs, etc. [28]. Fixed-wing UAVs and rotary-wing UAVs are the most common ones. Fixed-wing UAVs have the advantages of fast speed, long range, and large payload, but they require large landing sites and auxiliary equipment. Rotary-wing UAVs have the ability to hover and take off vertically. They are suitable for low-altitude and complex terrain flight but have shorter range and endurance. Rotary-wing UAVs are the widely used type of UAVs in military and civilian applications, and the common types include helicopter-type and multi-rotor-type [29]. Helicopter-type UAVs have large payloads and high flexibility, but the disadvantages are that the mechanical structure is more complex than multi-rotor, the operation difficulty is high, and the mainte-

nance cost is high. Multi-rotor UAVs, capable of vertical takeoff, landing, and hovering, are particularly suited for missions requiring low-altitude operations and stationary flight. At present, multi-rotor UAVs have become the main type of object detection research and are widely used in environmental protection, precision agriculture, and disaster rescue.

These UAVs develop fast because they are relatively cheap and have the ability to take pictures simply. UAVs can carry high-resolution cameras or other sensors, obtaining clearer and richer object information. Compared with orbital and other aerial sensing acquisition methods, UAV platforms can perform object detection in high-altitude or hazardous areas, avoiding the risk of casualties and cost expenditure. Therefore, UAV-based image acquisition systems are very popular in commercial and scientific exploration. However, visual inspection of objects by UAVs is still biased, inaccurate, and time-consuming. Currently, the real challenge regarding remote sensing methods is to automate the process of obtaining fast and accurate information from data, where detecting objects from UAV images is one of the key tasks.

UAV images for object detection are classified based on flight altitude and application areas at different altitudes: (1) eye-level view: this category, at altitudes between 0 and 5 m, is optimal for ground-level observations. (2) Low and medium height: 5–120 m. It represents the gap between most commercial and industrial applications. (3) Aerial imaging: higher than 120 m, this category is synonymous with high-altitude data capture and often requires special permissions.

2.2. UAV Sensors for Object Detection

UAV sensors are various devices that can measure the motion state, position information, and environmental parameters of UAVs. They are important components for achieving autonomous flight and mission execution. The use of UAVs depends on various factors, such as payload capacity, size, cost, safety, environment, redundancy level, and autonomy level. According to different measurement principles and functions, we will only introduce four sensor technologies that are important for UAV object detection:

Visual sensors: A visual sensor is a device that uses photoelectric sensors to obtain images of objects. From the image, state information such as location and speed of the object can be calculated. The more important thing for visual sensors is the processing algorithm. Recently, the development of deep learning algorithms has brought more extensive applications to visual sensors.

Ultrasonic sensor: Ultrasonic waves are sound waves that exceed the upper limit of human hearing frequency. Due to their good directivity and strong penetration, they are widely used for distance and speed measurement. The ultrasonic sensor emits a signal that is reflected by the object and then received by another ultrasonic sensor. An ultrasonic sensor is generally cheap, but its drawbacks include a low data update rate and a limited measurement range.

Laser sensor: The principle of the laser sensor is basically the same as the ultrasonic sensor, except for the different emitted signal. A laser source is emitted by the laser ranging sensor at the speed of light, which makes the signal frequency much higher than the ultrasonic sensor [30]. Its disadvantages are high price, small measurement range, and the ability to scan.

Ground-penetrating radar: Ground-penetrating radar (GPR) is a popular nondestructive testing technique for object detection and imaging in geological surveys [31]. Fast and accurate object detection on the surface can reduce computation time and hardware requirements.

Thermal imager: A thermal imager is a device that can convert invisible infrared radiation into visible images. It can detect the temperature distribution and thermal anomalies of objects in dark or harsh environments [32].

2.3. The Difference between UAV Object Detection and Common Object Detection

In normal view, the datasets used for object detection algorithms are mostly taken by handheld cameras or fixed positions, so most of the images are side views. However, UAV aerial images have different characteristics from ordinary view images because they are taken from a top-down view. This means that the object detection algorithms in normal view cannot be directly applied to UAV aerial view.

Firstly, the quality of UAV aerial images is affected by many factors, such as the instability of equipment causing jitter, blur, low resolution, light change, image distortion, etc. These problems need to be preprocessed for the video to improve the detection effect of methods [33].

Secondly, the object density in the aerial view is inconsistent and the size is very small. For example, pedestrians and cars may occupy many pixels in normal view but only a few pixels in aerial view, and they are distributed irregularly, causing object deformation, increasing the difficulty of multi-object detection, and requiring special network modules to extract features [34].

Finally, the occlusion in aerial view is also different from that in normal view. In normal view, the object may be occluded by other objects, such as a person in front of a car. However, in an aerial view, the object may be occluded by the environment, such as buildings and trees [35]. Therefore, it is not possible to directly apply the multi-object detection algorithms trained on normal view video datasets to UAV aerial images. It is necessary to design corresponding algorithms, which can meet the UAV object detection task requirements according to the features of UAV images.

2.4. Challenges in UAV Object Detection

The object detection task in UAV remote sensing images faces many challenges, such as object rotation, complex background, an increase in small object issues, low detection efficiency caused by scale changes, and sparse and uneven distribution of object categories. Detailed explanations of the different challenges are as follows:

Small objects increasing problem: The scale range of objects in UAV images is large. Buildings, pedestrians, mountains, and animals often appear in the same image. Small objects have a very small proportion in the image, which makes detection difficult [36,37]. The multiscale feature fusion method can effectively solve the problem of small objects increasing by detecting objects of different sizes through different levels of features.

Background complexity problem: The dense object areas in UAV images contain many identical items, which increases the probability of false detection. In addition, a large amount of noise information in the background of UAV images can also weaken or obscure the object, making it difficult to detect continuously and completely [38]. In order to improve detection accuracy and robustness in the complex background, attention mechanisms and graph neural networks can be used to enhance the relationships between objects.

Category imbalance problem: The objects in the images captured by UAVs may have category imbalance problems, such as having a large number of objects in one category and a small number of objects in the other category, resulting in the detector leaning towards predicting categories with a large number [39]. Generative adversarial networks or autoencoders can be used to enhance data diversity and quality, which will alleviate issues regarding data imbalance and noise.

Object rotation problem: In UAV images, objects can appear in any position and direction [40]. Traditional object detection algorithms usually assume that the object is horizontal, but, in UAV images, the object may be rotated at any angle. In addition, rotating objects may change their shape and appearance in the image, which makes the object detection algorithm based on shape and appearance not work accurately [41]. Use a rotation box or polygon box to represent the object that can adapt to any angle of the object.

3. UAV Object Detection Method Based on Deep Learning

With the development of UAV technology, UAVs equipped with cameras and embedded systems have been widely used in various fields, including agriculture [42], power inspection [43], aerial photography [44], and mapping [45]. These applications require the UAV platform to have the ability to perceive the environment, understand scenes, and make corresponding reactions. The most basic function is automatic and efficient object detection. Object detectors based on deep networks [46,47] extract image features automatically through convolutional neural networks, greatly improving the performance of object detection.

In this paper, we focused on researching papers about deep-learning-based UAV object detection that were published in the past five years. Firstly, all papers were classified based on one-stage and two-stage object detection, and then four common problems in UAV object detection were summarized. In addition, research has been conducted on applications such as geological environment surveys, traffic monitoring, and precision agriculture. Finally, classic UAV object detection algorithms based on deep learning targeting these common problems were selected [48]. In this section, we will describe the improvement methods proposed by various scholars based on the selection criteria above.

3.1. Object Detection Development Process

The development of object detection algorithms can be divided into two stages: traditional object detection algorithms and deep-learning-based object detection algorithms. Deep-learning-based object detection algorithms are further divided into two main technical routes: one-stage and two-stage algorithms [49]. Figure 1 shows the development of object detection from 2001 to 2023. Traditional object detection algorithms are mainly based on sliding window and artificial feature extraction methods, which generally consist of three steps: region proposal, feature extraction, and classification regression. The region proposal is to obtain the regions of interest where the object may be located. In the feature extraction process, artificial feature extraction methods are utilized to transform images in candidate regions into feature vectors. Finally, the classifier classifies objects according to the extracted features. These algorithms have the disadvantages of high computational complexity, weak feature representation ability, and difficulty in optimization. Representative algorithms include the Viola–Jones detector [50], the HOG pedestrian detector [51], etc.

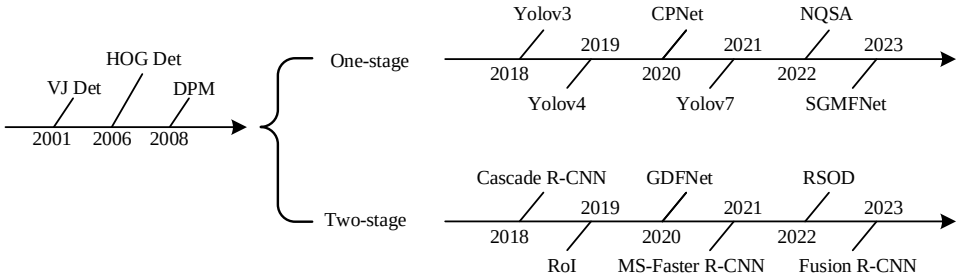

Figure 1. The development of object detection from 2001 to 2023.

In 2012, with the rise of neural networks (convolutional neural network, CNN), object detection developed fast. Deep neural networks are mainly used in deep-learning-based object detection algorithms to automatically extract high-level features from input images and classify objects. These algorithms have the advantages of fast speed, high accuracy, and strong robustness. In addition, the two-stage detector only processes the content of region proposals in the second stage, which results in the loss of position information of the object in the entire image. As shown in Figure 2, the two-stage object detection framework solves the problems of boundary object detection, inaccurate localization, and multiscale object. These algorithms generate proposal regions in the first stage and perform classification and regression on the contents in the regions of interest in the second stage. These algorithms

usually have high accuracy but low speed. Representative algorithms include R-CNN (Region-based CNN), SPP-Net [52], Fast R-CNN [53], Faster R-CNN, etc.

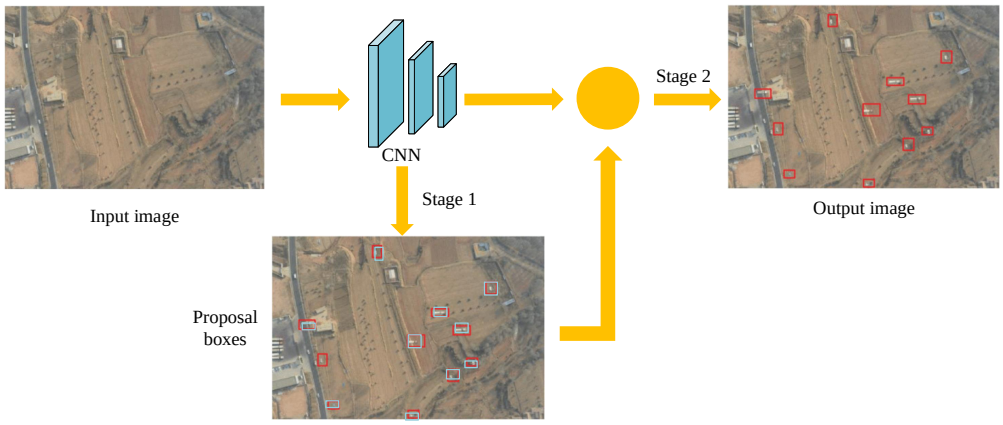

Figure 2. Two-stage object detection framework.

Due to the fact that two-stage detectors need two stages to perform detection, the first stage generates many proposal regions, resulting in large computation and low speed, which is not suitable for real-time scenarios. One-stage detectors solve these problems, and their framework is shown in Figure 3. One-stage object detection algorithms directly generate location and category from the image, removing the process of proposal region, thus having faster speed. However, one-stage object detection algorithms are susceptible to incorrect detection in localizing and detecting small objects due to the high number of densely generated region proposals. In addition, the classification and regression branches of one-stage methods are usually simple and difficult to capture the detailed features of the target, leading to unstable detection performance. Representative algorithms include SSD, Yolo series, etc.

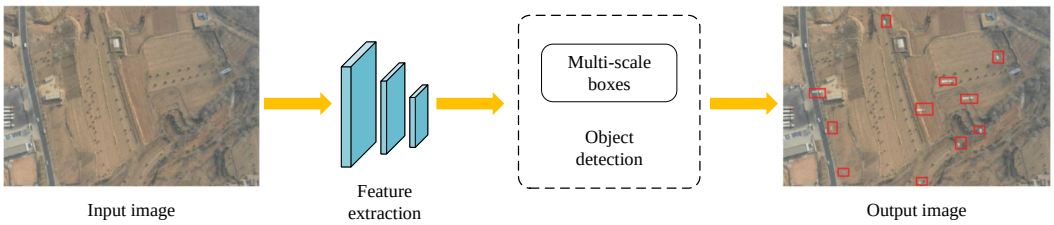

Figure 3. One-stage object detection framework.

3.2. One-Stage UAV Object Detection Algorithm

One-stage UAV object detection is a method that uses deep learning techniques to directly predict the location and category of UAVs from images. One-stage object detection algorithms only need to process the image once to obtain the classification and location information of the object. Thus, one-stage object detection algorithms have fast speed, which can be applied to scenarios with high real-time requirements. In 2015, Redmonn et al. proposed the Yolo algorithm. It divides the input image into fixed-size grids, and each grid predicts a certain number of bounding boxes, confidence scores, and probabilities for the category of objects. In 2016, Liu et al. proposed the SSD algorithm. SSD algorithm generates a set of default bounding boxes by using convolutional filters on different scales of feature maps and predicts the category and position offset of the object in the box. Figure 4 shows the comparison of Yolo and SSD algorithm structures.

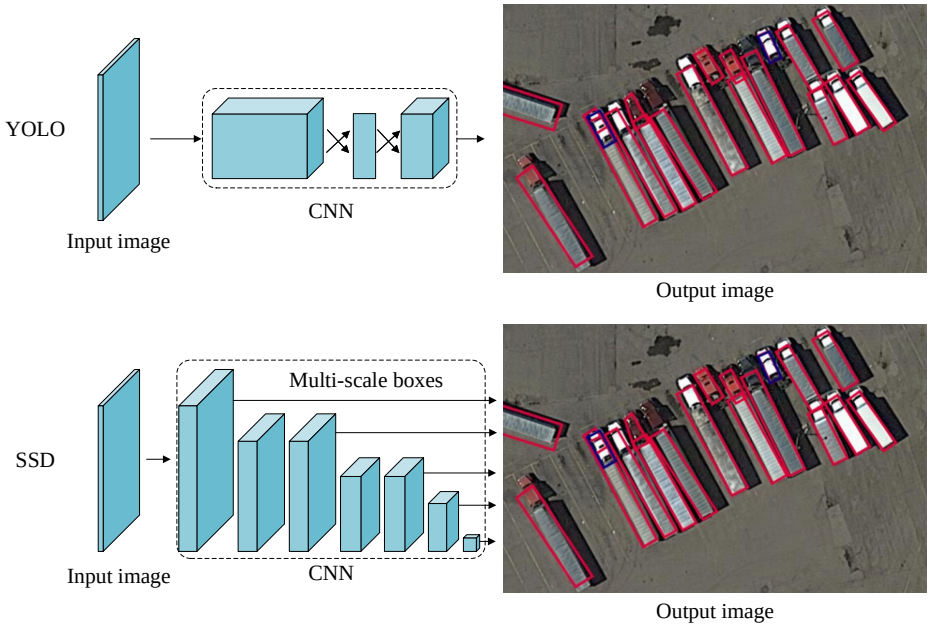

Figure 4. The comparison between Yolo and SSD algorithm.

Due to their fast running speed and high detection accuracy, one-stage detectors Yolo and SSD are widely applied. For example, Hossain et al. [54] transferred Yolo and SSD to the embedded device GPU JetsonTX2 to achieve UAV ground object detection and tracking. Lu et al. [55] integrated Yolov5 with shallow feature information, effectively improving the efficiency of UAV marine fishery law enforcement. Marta [56] proposed a method using dense point clouds to estimate obstacle heights and used the Yolo algorithm to detect atypical aviation obstacles. This method achieved the identification and classification of atypical aviation obstacles. Li et al. [57] integrated SSD with a convolutional block attention mechanism (SSD-CBAM) for earthquake disaster building detection.

In addition to directly applying the original one-stage detectors to multi-object detection under the UAV perspective, many scholars have improved them from various aspects such as network structure optimization, multi-task learning, introducing region-specific contextual information, and integrating multiple networks. Table 1 shows the comparison of object detection algorithms based on one-stage. The following will mainly introduce the object rotation problem, complex background problem, small object problem, and class imbalance problem in UAV one-stage detection.

Table 1. Comparison of main object detection algorithms based on one-stage method.

Problems	Models	Dataset Used	Published Year	Category
Small objects	Sommer et al. [58]	DLR 3K [59]	2017	Vehicle
	UAV-Yolo [9]	UAV-viewed	2020	Pedestrian
	SGMFNet [60]	VISDRONE2019	2023	Multicategory
Complex background	Segment-before-detect [61]	VEDAI [62]	2017	Vehicle
	Li et al. [63]	DOTA-v1.5 [64]	2020	Multicategory
	NQSA [65]	IR-UAV	2022	Infrared image
Category imbalance	CPNet [66]	xView [66]	2020	Pedestrian
	DS Yolov3 [67]	UAVDT [68]	2021	Multicategory
Object rotation	FS-SSD [69]	PASCAL VOC [70]	2020	Multicategory
	RSSO [71]	RSSO	2022	Multicategory

3.2.1. Aiming at Small Object Problems

The scale range of objects in UAV images is large. Buildings, pedestrians, mountains, and animals often appear in the same image. Small objects occupy a very small proportion of the image with limited resolution, which makes detection difficult.

In the early research of UAV object detection, Sevo and Avramovic [72] proved that CNNs can be effectively integrated into the object detection algorithm of aerial images. Sommer et al. [58] applied Fast R-CNN and Faster R-CNN to solve the problem of vehicle detection from aerial images. Although CNNs have a certain generalization ability, small objects occupy very few pixels in the image and it is difficult to extract effective features. In addition, CNNs are not robust enough to the rotation, occlusion, illumination, and other changes regarding small objects, which can easily cause false positives or false negatives.

To improve the performance of small object detection, Liu et al. [9] proposed UAV-Yolo, which optimized the Resblock in darknet by connecting two ResNet units with the same width and height. Zeng et al. [73] integrated a hybrid attention mechanism with coordinate-related attention and a multi-layer feature structure fusion, which can effectively distinguish the foreground and background features of aerial images and enrich the semantic information of shallow features. The algorithm performed well on the VisDrone2020 dataset, improving accuracy while ensuring detection speed. Qian et al. [74] combined Haar-Like features and MobileNet-SSD algorithm, using a top-down and horizontal connection method to construct a feature pyramid with high resolution and strong semantics, thus achieving multiscale UAV feature representation and detection. Tian et al. [75] proposed a DNOD method that used the VGG network to extract feature maps of UAV images and combined them with the location information of suspected regions for secondary identification, reducing the false negative rate of small objects. They combined Yolov4 and EfficientDet-D7, respectively, to verify the reliability and effectiveness of the algorithm.

Although these methods can detect small objects, their detection performance is poor when dealing with images containing a large number of dense small objects. To solve these problems, Zhang et al. [60] proposed SGMFNet, a network for identifying objects in aerial images captured by UAVs. SGMFNet uses self-attention guidance and multiscale feature fusion to improve detection accuracy. The network structure of SGMFNet is shown in Figure 5.

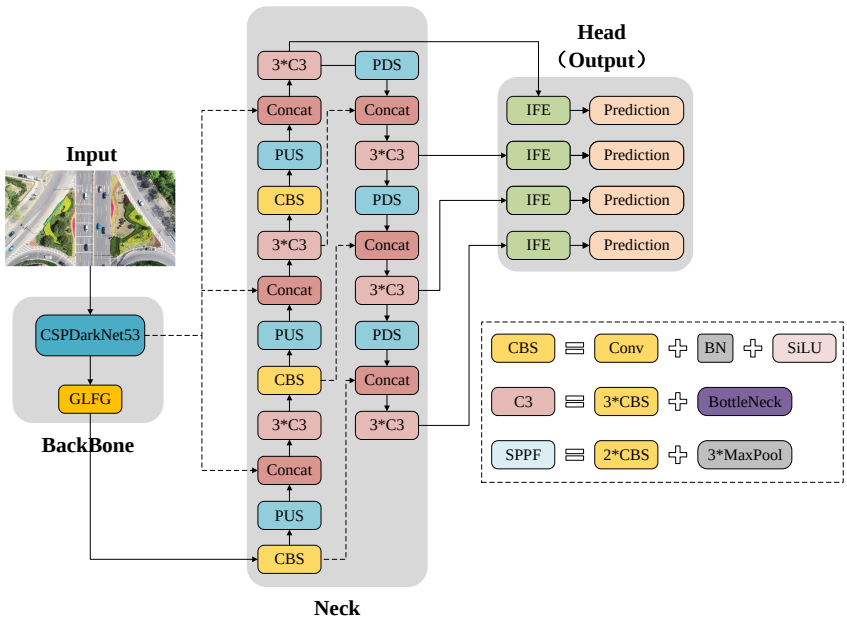

Figure 5. The network structure of SGMFNet.

The object detection process of SGMFNet has three steps. First, SGMFNet uses CSP-DarkNet53 to obtain features from the image. Then, it adds a GLFG module to the backbone

network. This module helps to combine local and global information. By applying self-attention, the GLFG module can measure the global resemblance of feature maps and thus obtain global information. Self-attention mechanism uses three matrices to calculate the attention weights of each element to other elements in the sequence and generates a new sequence representation according to these weights. Self-attention calculation is as follows:

$$Q = XW^q, K = XW^k, V = XW^v \tag{1}$$

where X denotes the input sequence. W^q, W^k, and W^v are three random initialization matrices, which can be adjusted by attention dynamically:

$$Attention(Q, K, V) = \text{Softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}})V \tag{2}$$

where the global similarity is calculated by $Q \cdot K^T$; d_k is the number of channels in the hidden layer; Softmax normalizes the result of the calculation to obtain a weight coefficient of V. Attention mechanism can be obtained by multiplying the results of the computation.

Next, the neck is composed of the PSFF module, which fuses multiscale features from different sampling branches that run in parallel and capture various features.

Finally, in the head, the shallow feature map is enhanced by the IFE module to boost the detection accuracy of small objects.

SGMFNet's experimental effects for small objects were evaluated on the VISDRONE2019 dataset. The experimental results are shown in Table 2 and the qualitative results are shown in Figure 6.

Figure 6. The qualitative results of SGMFNet: (**a**) Yolov7; (**b**) mSODANet; (**c**) SGMFNet. Yellow wireframes highlight the areas that exhibit the superiority of SGMFNet approach.

Table 2. The results of SGMFNet.

Method	mAP_{50} (%)	mAP_{75} (%)	mAP (%)	AP-Small (%)	AP-Mid (%)	AP-Large (%)
SGMFNet-m	62.3	40.2	39.5	31.9	49.4	52.1

Note: mAP denotes the mean average precision; AP means the average precision.

As shown in Figure 6, in the first column of images, complex backgrounds occupy an important part of the image. The second column of images was captured from a low altitude angle by a UAV, and there are significant differences in the scale of the objects due to distance. The third column image contains a large number of dense small objects. The observation results indicate that the SGMFNet method exhibits significant advantages over other methods. The experimental results of Table 2 and Figure 6 show that SGMFNet has significant advantages over the other methods in processing images containing a large number of dense small objects.

3.2.2. Aiming at Complex Background Problems

In UAV images, target-dense areas and background noise information can interfere with object detection and lead to false alarms. This not only reduces the performance and safety of UAVs but also endangers the stability and development of society.

Although CNNs have generalization ability, convolution and pooling operations will lose details of feature maps, which is not conducive to complex background object detection. To solve these problems, some effective improvement algorithms have been proposed. Using a deep fully convolutional network, Audebert et al. [61] segmented vehicles in aerial images with high accuracy and detected them by finding connected components. By combining semantic segmentation and object detection in aerial images, they demonstrated that detection performance can be enhanced, especially in obtaining object boundary information. Li et al. [63] constructed a semantic segmentation-guided RPN (sRPN) module to suppress background clutter in aerial images. This module integrates multi-layer pyramid features into a new feature after performing Atrous spatial pyramid pooling (ASPP) and convolution operations. Inspired by this, Chen et al. [76] combined a spatial pyramid pooling network with a probabilistic pooling method, enhancing the robustness of the aerial image recognition algorithm. They solved the problem of invalid features of complex backgrounds having a negative impact on recognition accuracy and achieved higher aerial image recognition accuracy.

The attention mechanism is a technique inspired by human perception that can filter out useful information for tasks. For example, Zhang [65] transformed local contrast measurement into non-local orthogonal contrast measurement in deep feature space. In addition, they regarded feature points that disrupt semantic continuity as potential object locations by the self-attention method. On this basis, they designed a multiscale one-stage detector by effective receptive field calculation. The network structure is shown in Figure 7. By using a bidirectional serial feature modulation method, they can fully retain the multiscale features of the object. This method ensures the accuracy of object detection in complex backgrounds and meets the real-time requirements. Figure 8 shows the representative results based on NQSA from the MWIR-UAV dataset.

The results show that the detection layer can learn with the bidirectional high-level feature modulation guided by low-level features, and better object detection performance can be achieved.

In UAV images, contextual information can help the attention mechanism distinguish between object and background, thus improving the quality of object features. For example, Chen et al. [77] proposed a context-based feature fusion method based on deep CNN, which used different levels of feature maps for fusion, improving the matching degree between region proposal and object features. Yue et al. [78] proposed a global–local context collector, which extracted global and local contextual information and enhanced low-quality objects in complex backgrounds.

The existing deep-learning-based UAV object detection methods in complex backgrounds still have defects. The attention mechanism can effectively filter out useless information in the background, but, in the specific scenario of UAV object detection, it needs to reasonably allocate weights to avoid false detection of small objects. Contextual information can improve the model's understanding of background and foreground,

but contextual information needs to be filtered. Therefore, it is of great significance to study how to detect objects in complex backgrounds.

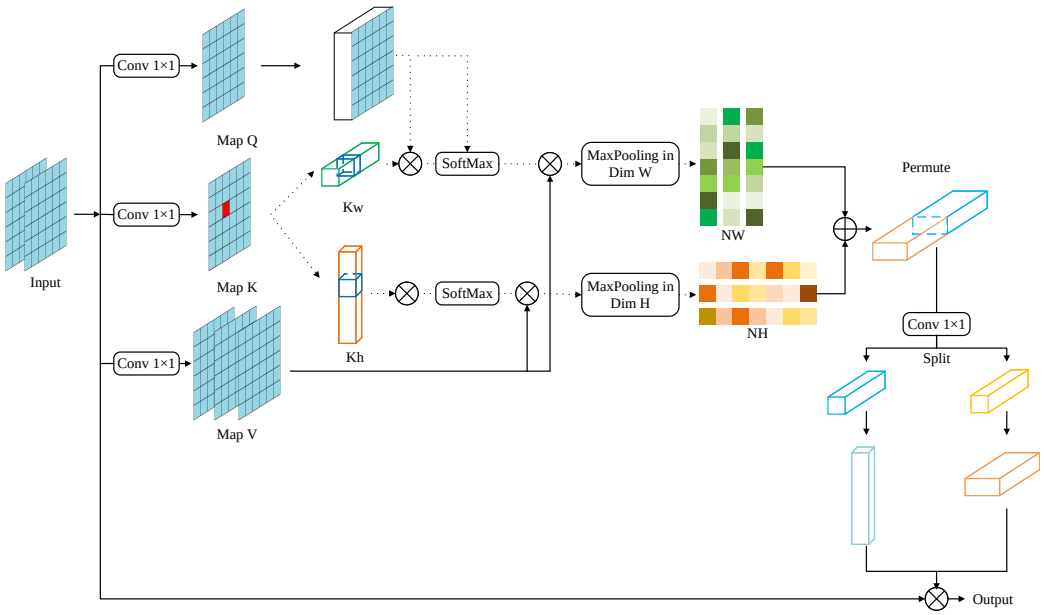

Figure 7. The structure of NQSA model.

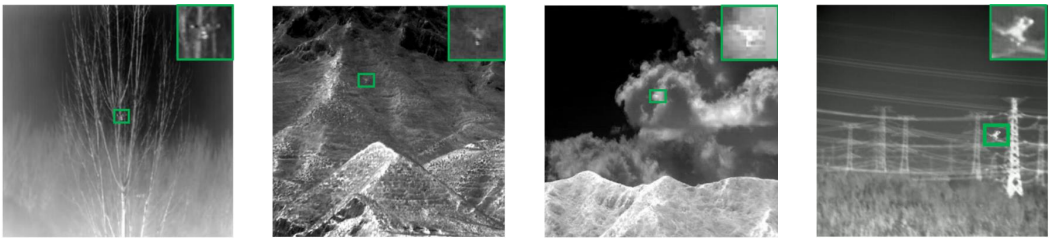

Figure 8. The results of NQSA from the MWIR-UAV dataset. Green wireframes represent schematic of typical objects.

3.2.3. Aiming at Category Imbalance Problems

UAVs can cover a wide area without being limited by terrain, so the images of UAVs have a large field of view. In these images, some categories of objects rarely appear, while some categories of objects appear frequently, resulting in low efficiency and accuracy in object detection.

To deal with this problem, Li et al. [67] proposed DS Yolov3, which detects objects of different scales through a multiscale perception decision discrimination network, and designed a multiscale fusion channel attention model. However, directly using convolutional detectors to process these images is too costly. Although the sliding window method can crop the images, it is not efficient. So, many algorithms improve efficiency by reducing the search area. Yang et al. [79] proposed an aerial image object detection algorithm based on object clustering for the problem of imbalanced object distribution in aerial images. The algorithm unified object clustering and detection in an end-to-end framework. Ju et al. [80] proposed an aerial image object detection algorithm based on the density map of the object center point, which mainly consists of three modules: density map generation module, region segmentation module, and detection fusion module. The distribution of pixel intensity in the whole image can be well-reflected by the density map. The region segmentation module crops out candidate detection regions based on the information from the density map. Finally, the region detection results and original images are fused.

Considering the shortcomings of candidate region generation algorithms, some studies apply reinforcement learning to object search in large field-of-view images. For example, Burak [66] proposed an efficient object detection framework for large field-of-view visible light remote sensing images, consisting of two modules called coarse-level search and fine-level search, as shown in Figure 9. Both searches are cascaded algorithms that combine reinforcement learning with CNN.

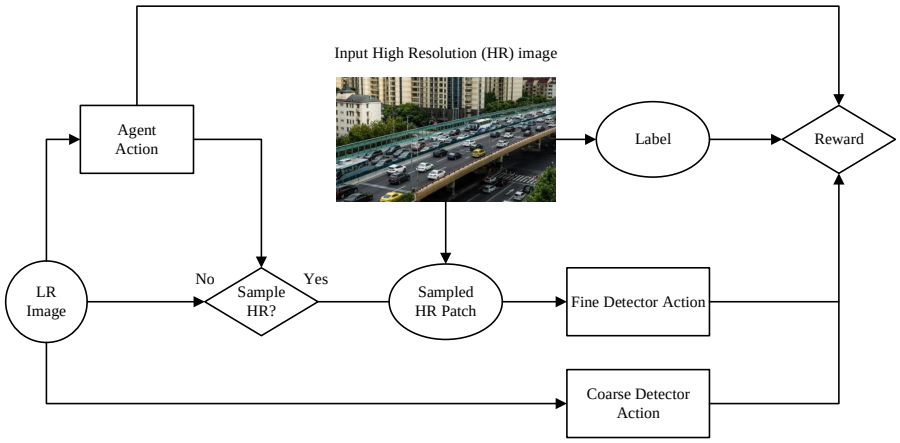

Figure 9. Bayesian decision influence diagram. At training time, HR (high-resolution) images are downsampled to LR (low-resolution) images. Detector uses LR/HR image to output bounding boxes.

In Figure 9, circles represent random variables, squares/rectangles represent actions, and diamonds represent utilities. In the coarse search, the low-resolution image is divided into sub-images of the same size and their respective rewards after magnification are calculated. In the fine search, the sub-images selected by the coarse search module are further optimized in the search space to finally determine which sub-images to magnify. Detection results are shown in Figure 10 and Table 3.

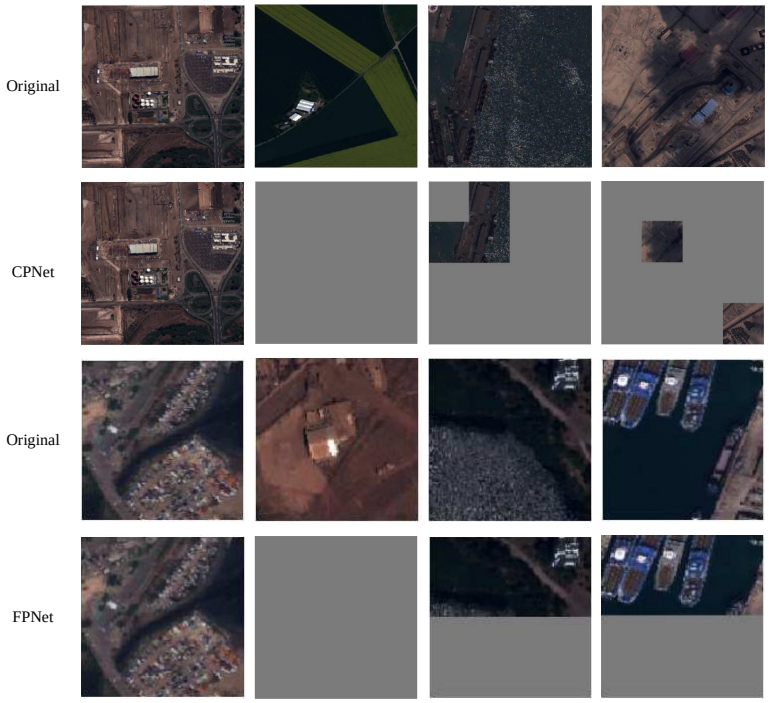

Figure 10. The learned policies by the coarse- and fine-level policy. The top row shows the initial LR images provided to the coarse policy network. The second row shows the patches selected by the coarse policy network, while the last two rows represent the policy learned by the fine-level policy network. The regions for using LR images are shown in gray.

Table 3. The detection results of CPNet method.

Model	Coarse Level			Fine Level			Coarse+Fine Level		
	AP	AR	R-HR	AP	AR	R-HR	AP	AR	R-HR
CPNet	38.2	59.8	40.6	38.3	59.6	35.5	38.1	59.7	31.5

Note: AR denotes average recall. R-HR means the ratio of sampled HR image.

Tests conducted on the xView dataset demonstrate that the CPNet retains nearly the same accuracy as the object detector that exclusively uses HR images with a running time of only 30% of HR images.

There are other improved methods based on deep learning used in the UAV object detection for category imbalance problems. For example, Wang et al. [81] proposed a reinforcement-learning-based method for UAV object searching and tracking, which enables UAVs to autonomously search and track moving objects in complex environments while considering UAV motion constraints and energy consumption. Yudin et al. [82] presented an object detection algorithm in aerial images based on the density map of the object center point, which can quickly and accurately identify the position and direction of all vehicles in the intersection. Then, the results of object detection are used as additional features input to two modern efficient reinforcement learning methods (Soft Actor–Critic and Rainbow), which can accelerate the learning process of agents.

In summary, UAVs need to monitor and collect large areas of space or scenes in fields such as pedestrian detection, remote sensing mapping, agricultural monitoring, etc. Fast and accurate searching and detection of objects in large field-of-view images can effectively save computing time, reduce hardware requirements, and improve UAV work efficiency.

3.2.4. Aiming at Object Rotation Problems

Objects in UAV images may appear at any position and orientation, and the angular variation in objects of the same class is also different. Rotating objects can lead to a high percentage of false or missed detections due to erroneous and unstable locations.

To solve this problem, some researchers have proposed rotation-based object detection methods, which represent the position and orientation of objects by predicting their center point, width, height, and rotation angle. This method can better adapt to the rotation variation in objects and improve the detection accuracy and robustness. For example, to address the issue of rotating object recognition accuracy in UAV optical remote sensing images, Wang et al. [71] developed a rotating object detection technique for remote sensing images that is based on feature alignment of candidate areas (RSSO). The paradigm for RSSO object detection is displayed in Figure 11.

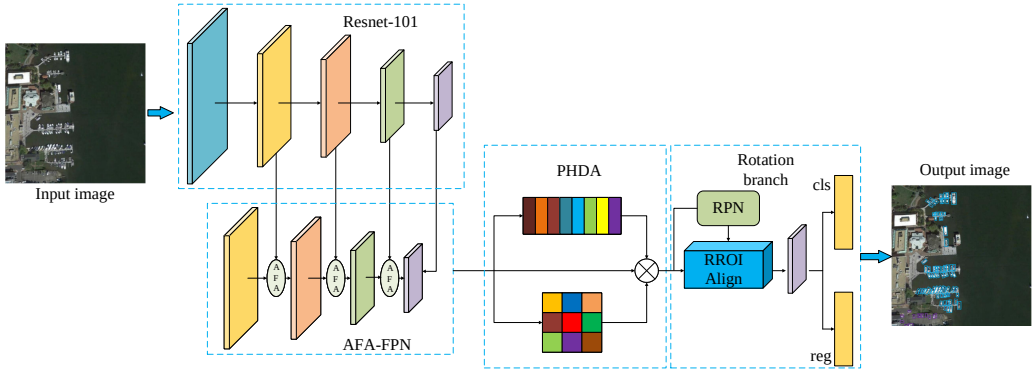

Figure 11. The pipelines of the RSSO model.

The RSSO network uses ResNet-101 [83] as the backbone for feature extraction. Then, the network adjusts the relationship between feature maps through the AFA-FPN module, solving the problem of inaccurate feature alignment between adjacent layers and ensuring the deep semantic features of small objects can be effectively transferred to the shallow

feature map. The PHDA module combines parallel channel domain attention and spatial domain attention, identifies the local regions in the feature map tensor that contain small object features, and assigns larger weights to these regions to suppress the influence of background noise. The rotation branch uses rotation regression, obtains rotation-invariant area features, minimizes the mismatch between region features and objects, and reduces the regression difficulty of the rotation candidate boxes. The results of the experiments are shown in Figure 12 and Table 4.

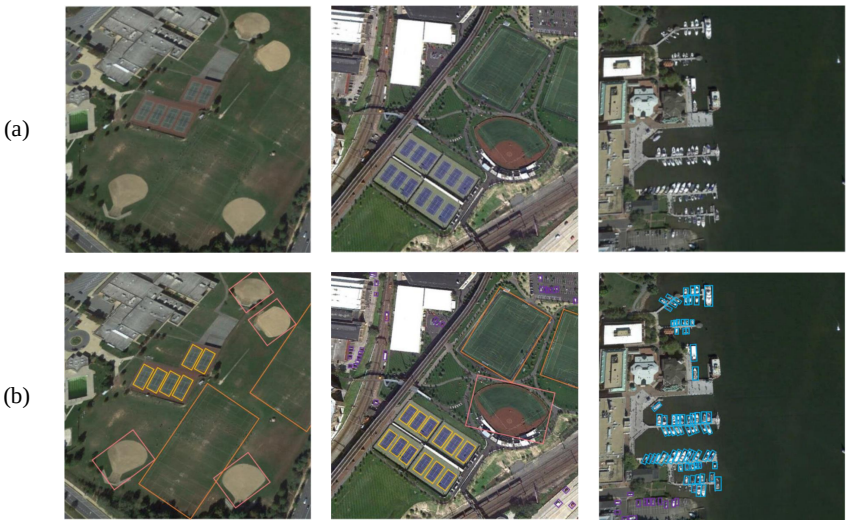

Figure 12. The experimental results of the RSSO method: (**a**) original images; (**b**) detection results. The first column shows the detection results of a single type of object in the simple background. The second column shows the detection results for multiscale objects on high-resolution remote sensing images of large scenes. The third column shows the detection results of densely distributed small objects with different angles.

Table 4. The results of RSSO method.

Method	PL	BD	ST	SH	TC	BR	RA	HC	mAP
RSSO	82.43	79.46	82.21	79.02	89.94	77.63	88.31	84.81	82.04

Note: PL, BD, ST, SH, TC, BR, RA, HC are the object categories for Plane, Baseball diamond, Storage tank, Ship, Tennis court, Bridge, Roundabout, and Helicopter, respectively.

The results in Figure 12 and Table 4 show that the method has high accuracy and recall rate in UAV remote sensing image object detection. However, rotation-based methods also have some drawbacks, such as large computational cost, high training difficulty, unstable angle regression, etc. Therefore, how to effectively deal with rotating objects in UAV images is still a worthy research problem.

In this field, there are some other important pursuits. For example, to detect rotating objects in UAV images, Liang et al. [69] developed a one-stage detector with spatial context analysis based on feature fusion and scale modification. Xiao et al. [40] first used CNN and rotation anchor mechanism to generate candidate rotation regions, which represent the position and orientation of objects. Then, using bilinear interpolation and rotation pooling operations, the irregular rotation regions are transformed into regular rectangular regions, which are used to extract the features of objects. Finally, using a fully connected layer and Softmax layer, each rotation region is classified and regressed, outputting the category and position of the object.

3.3. The Two-Stage Object Detection Algorithm

Two-stage UAV object detection is a method that performs classification to determine categories after proposing the regions of interest (ROI). The two-stage method has higher

accuracy than the one-stage method because the region of interest is located and classified by the two-stage detector from the first stage. However, the two-stage method takes more time to infer than the one-stage method due to extra regions and stage processing.

In 2014, Girshick et al. [11] tried to combine the Region Proposal and CNN based on AlexNet and proposed R-CNN algorithm with greatly improved detection performance. He et al. [52] used the Spatial Pyramid Pooling (SPP) module in CNN, which solves the limitation of fixed-size images and avoids repeated extraction of image features.

To solve the problem that the R-CNN algorithm has a great deal of redundancy and runs slowly in the feature extraction operation, Grishick [53] proposed a Fast R-CNN using region of interest (ROI) pooling based on R-CNN and SPP-Net algorithm structure to achieve end-to-end detection. Ren et al. [12] presented the Faster R-CNN method that replaces the selective search algorithm with the region proposal network (RPN) for generating candidate regions more efficiently. The network further improves the detection speed by sharing the convolutional features.

Table 5 shows the comparison of object detection algorithms based on two-stage. The following will mainly introduce the object rotation problem, complex background problem, small object problem, and category imbalance problem in UAV two-stage detection.

Table 5. Comparison of main object detection algorithms based on two-stage detection.

Problems	Reference	Dataset Used	Published Year	Category
Small objects	R-DFPN [84]	Google Earth	2018	Ship
	RSOD [85]	UAVDT [68]	2022	Traffic
Complex background	SCRDet [86]	DOTA [64]	2019	Multicategory
	Shao et al. [87]	UAV-head [87]	2021	Pedestrian
	FR-Transformer [88]	UWHD [88]	2022	Agriculture
Category imbalance	Deng et al. [14]	NWPU VHR-10 [89]	2018	Multicategory
	MLD [90]	Leaf [90]	2022	Agriculture
Object rotation	RoI Transformer [91]	DOTA [64]	2019	Multicategory
	TS^4Net [92]	UAV-ROD [92]	2022	Vehicle

3.3.1. Aiming at Small Object Problems

The object scale in UAV images varies greatly and the proportion of small objects is high. Therefore, to perform a fast and accurate object detection, the cost of object detection search needs to be reduced first. Avola et al. [93] constructed a multi-stream structure to simulate multiscale image analysis. This structure limits the number of region proposals at the expense of the precision and reduces the number of parameters, which can detect objects in UAV video sequences in real time.

To improve the detection of small objects at different scales, Lin et al. [94] presented a feature pyramid network (FPN) that combines low-level features with high-level semantics to enhance the detection. The FPN builds a pyramid of features using different levels of CNN feature maps, and then predicts independently on each layer of the pyramid. This model can greatly reduce the computational cost and solve the problem of inconsistency between training time and testing time.

Using the FPN algorithm, Yang et al. [95] incorporated the dense connection from DenseNet to obtain features with higher resolution. They used lateral and dense connections in the top-down network to combine features from different levels. These algorithms improved the effect of small object detection to some extent, but new modules increased the computational cost, and the speed of the algorithm was difficult to guarantee. Liu et al. [96] proposed a high-resolution detection network HRDNet, which used multi-resolution input and had multiple depth backbones. At the same time, they designed a Multi-Depth Image Pyramid Network (MD-IPN) and a Multi-Scale Feature Pyramid Network (MS-FPN). MD-IPN used multiple depth backbones to maintain multiple position information, extracting various features from high resolution to low resolution, solving the problem of small object

context information loss, and significantly improving the detection speed compared with FPN. A high-resolution feature extractor that assigned features to different levels was designed in [97]. In [97], the authors applied a coarse-to-fine region proposal network to gradually raise the IoU threshold intersection, and maintained sufficient positive anchors at each stage. To more accurately estimate the mask, they scored the mask head with both classification score and location score.

However, during CNN downsampling, important features of an object are degraded. In addition, the object observed by UAV has the characteristics of small size and high density, which increase the difficulty of object detection. To solve above problems, Sun et al. [85] proposed a real-time small object detection algorithm for identifying small objects in the air called RSOD. This algorithm is specifically designed for monitoring traffic with UAVs. The algorithm is based on the Faster R-CNN framework, and introduces a multiscale feature fusion module and an adaptive anchor box generation module to improve the feature expression and localization ability of small objects. The structure of the RSOD model is shown in Figure 13. First, the image is input into the backbone network. Then, information fusion is performed on four feature layers with different depths. After that, the adaptive weighted feature fused four features. Finally, the object classes are predicted by the Yolo Head module. The experimental results are listed in Table 6, and some detection results based on the RSOD network are shown in Figure 14.

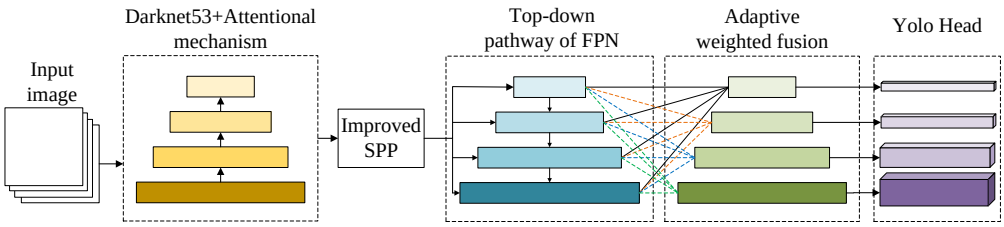

Figure 13. The overall structure of the RSOD model.

Figure 14. Thedetection results of the RSOD model.

Table 6. The object detection results (mAP_{50} (%)) based on the RSOD model.

Method	Input Size	Car	Bus	Truck	All
RSOD	416	41.5	41.5	43.2	43.9
	608	57.3	49.8	51.1	52.7

The results of the RSOD model achieved 29.8% mAP on the VisDrone-2019 dataset while maintaining a high detection speed.

Small object detection algorithms have a wide range of application needs in the fields of face detection, military reconnaissance, and traffic security. In the autonomous flight of UAVs, the identification and spatial localization of ground landing points at higher flight altitudes and also has high requirements for the accuracy of small object detection.

3.3.2. Aiming at Complex Background Problems

Object detection has made significant progress, but it still faces difficulties with objects that have arbitrary orientation and dense distribution. These problems are especially prominent for aerial images with complex backgrounds. For UAV object detection in arbitrary orientations, Xu et al. [98] added four sliding offset variables to the classic horizontal bounding box representation to achieve horizontal detection of objects in complex backgrounds and oriented detection of other objects. Zhang et al. [99] proposed a GDFNet (Global Density Fused Convolutional Network) model, which enabled object detection in crowded scenes with high distribution density.

Attention mechanism is a common technique for object detection as it helps to reduce the noise from complex backgrounds. For example, Yang et al. [86] introduced attention mechanism into object detection and proposed SCRDet, where a supervised Multi-Dimensional Attention Leaner (MDA-NET) was used to emphasize the features of the objects and reduce the features of the background, which can successfully detect the objects in complex scenes. Liu et al. [100] designed a center–boundary dual hybrid attention module that can boost ship detection accuracy. Shao et al. [87] presented an enhanced spatial attention module, which increased the precision of pedestrian detection in UAV images.

However, traditional attention mechanisms are typically built on CNN architectures and only calculate the weights of certain regions. To better capture more global context information and improve the accuracy of UAV image object detection, Zhu et al. [88] combined transformer with convolutional kernel attention mechanism and proposed the FR-transformer method to achieve the requirement of accurate and fast detection for wheat heads by UAVs. Figure 15 shows the network structure of FR-transformer. In this FR-transformer model, the traditional CNNs treat images as matrices or grids, combining neighborhood pixels or feature pixels through sliding windows; visual transformers divide the input image into multiple image blocks to form sequences and process sequential relationships using fully connected layers or attention mechanisms.

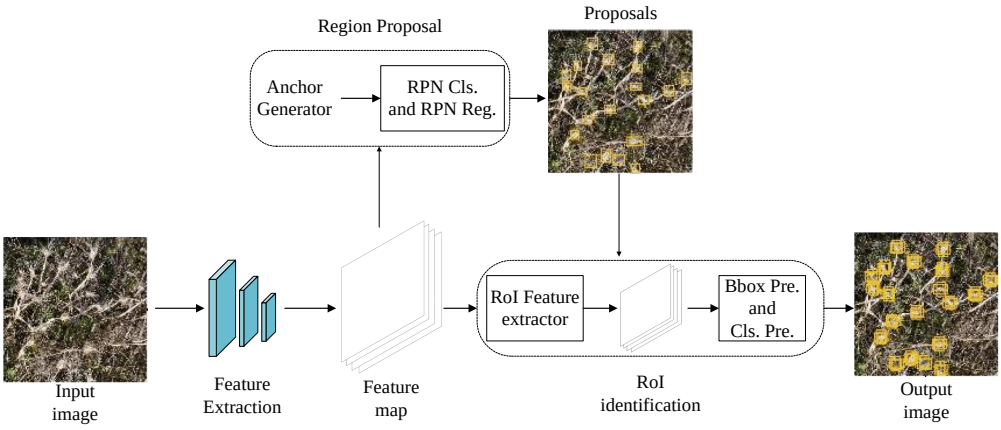

Figure 15. The structure of the FR-transformer method.

From the transformer architecture, the image (H, W) is first input to the Patch Partition for block operation and then sent to the linear embedding module to adjust the channel number. Finally, through multi-stage feature extraction and downsampling, the final prediction result is obtained. After each stage, the size shrinks to half of the initial size, and the channel grows to double of the initial channel, which resembles the ResNet network. Block in Swin transformer is used by the transformer block in each stage.

Figure 16 shows the qualitative results of the method in [88] and Table 7 shows the detection performance of the FR-transformer method. According to the results of the wheat detection task, the FR-transformer method can detect wheat heads with high accuracy in the real field environment. These research results show that UAV object detection is

very useful for agriculture. For example, UAV can be applied to agricultural inspections, providing high-resolution aerial images to help farmers accurately observe the condition of farmland. UAVs can also be used to help farmers monitor the environmental conditions of the land and better manage agriculture.

Table 7. The detection performance of the FR-transformer method.

Method	mAP_{50}	mAP_{75}	mAP	AP-Small	AP-Mid	AP-Large
FR-transformer	88.3	38.5	43.7	6.4	44	54.1

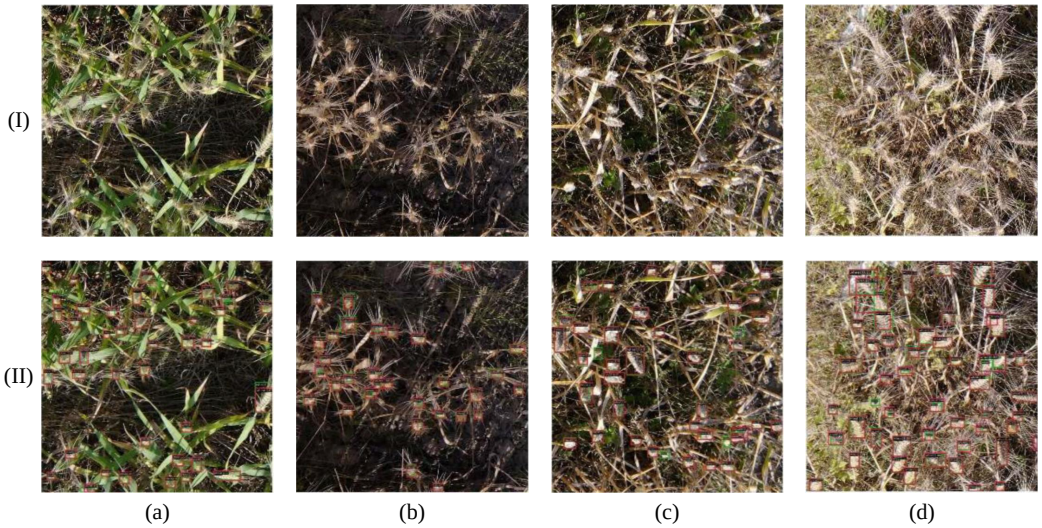

Figure 16. The detection results of FR-transformer: (I) RGB image; (II) FR-transformer. The challenges are (**a**) different growth stages; (**b**) brightness changes; (**c**) image blur; (**d**) overlap. The green bounding box represents the ground truth of the wheat, and the red bounding box represents the predicted area of the wheat.

As introduced above, the attention mechanism strengthens the feature extraction ability by distributing the feature information among the learning channels with different weights. However, it is still worth studying how to use the attention mechanism reasonably and control the size of the model.

3.3.3. Aiming at Category Imbalance Problems

In UAV images, the objects are generally sparse and non-uniform distribution, which makes the detection efficiency very inefficient. In addition, the two-stage detector only processes the content of the candidate region in the second stage, and the position information of a small number of categories in the whole image is missing. To address these problems, Deng et al. [14] proposed a detection method that is effective for multi-class objects in variable remote sensing images of large scale. Firstly, the feature extractor has the Relu and inception modules to make the receptive field size more diverse. Then, an accurate object detection network and a multiscale object proposal network are combined, which contain several feature maps, enabling small and densely packed objects to produce stronger response. Cores et al. [101] presented a two-stage spatiotemporal object detection approach that addresses the imbalance problem by considering temporal information. First, a short-term proposal linking and aggregation method was used to improve features. Then, a long-term attention module was designed to improve the short-term aggregated features with long-term spatiotemporal information. This module uses long-term connections between proposals in frames to deal with the imbalance of categories.

It is essential to estimate seed performance and yield by computing the amount of maize leaves from UAV images. However, detecting and counting maize leaves is difficult because of a variety of plant disturbances and the cross-covering of the adjacent seedling leaves. Thus, Xu et al. [90] proposed a method of detecting and counting maize leaves based on UAV images. To reduce the effect of weeds on leaf counting, maize seedlings were separated from the complex background by R-CNN technique. A new R-CNN loss function SmoothLR is proposed. Then, Yolov5 was used to detect and count the segmented leaves of maize seedlings. The flowchart of maize leaves method (MLD) presented in [90] is shown in Figure 17. The results of MLD method on the detection of maize leaves are shown in Figure 18 and Table 8.

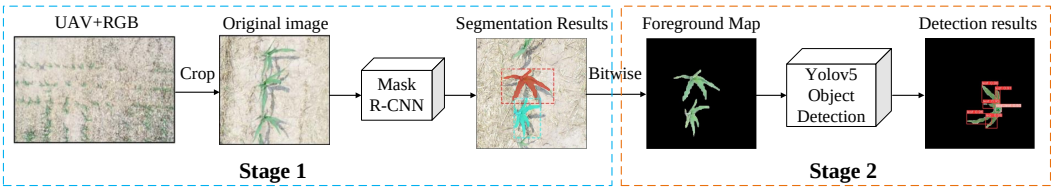

Figure 17. Flowchart of maize seedlings and leaves detection.

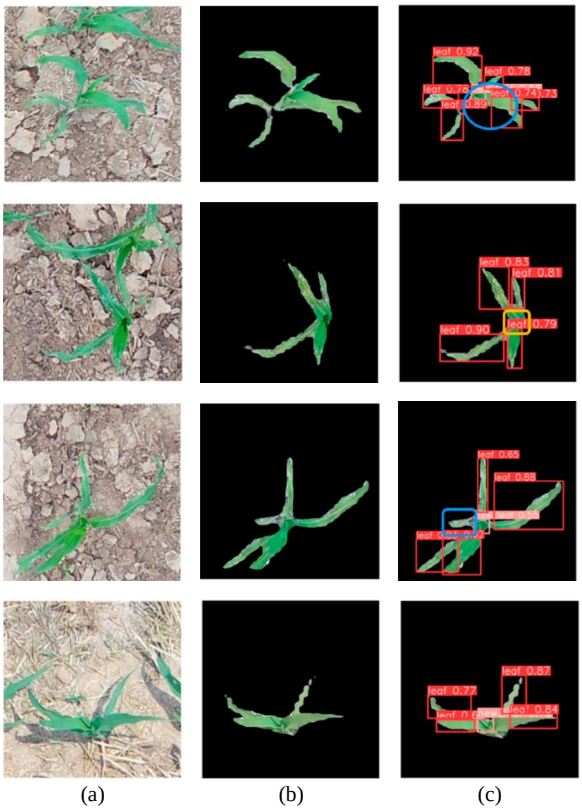

Figure 18. Visualization results of leaf detection: (**a**) original; (**b**) foreground map; (**c**) MLD results. Newly appeared leaves are in the red rectangular boxes. The blue circle is the multi-inspection of leaves. The blue and yellow rectangles are missing fully unfolded leaf and newly appeared leaf inspection.

From Table 8, we can see that the MLD method exhibited higher recall and AP than Faster R-CNN. The MLD method can accurately detect both fully unfolded leaves and newly appeared leaves in the image (see Figure 18).

Methods that introduce other networks are able to select different structures for different scene characteristics of object detection in UAV images. However, such methods have poor migration ability and generalization when dealing with tasks containing multiple scenes.

Table 8. Test results of Fully Unfolded Leaf (FUL) and Newly Appeared Leaf (NAL) detection.

Model	Precision (%)		Recall (%)		AP (%)		mAP (%)
	FUL	NAL	FUL	NAL	FUL	NAL	
Faster R-CNN	40.7	18.2	79.9	48.1	67.3	17.3	42.3
MLD	92	68.8	84.4	50	89.6	54	71.8

3.3.4. Aiming at Object Rotation Problems

The rotation of the object makes its position in the image become uncertain, which increases the complexity of position regression, especially for high-resolution aerial images. Thus, Ding et al. [91] proposed a module called RoI Transformer to address the problem of detection interference caused by changes in the angle of view and rotation of the object from the UAV perspective, with the aim of achieving the detection of oriented and dense objects through supervised RRoI learning and location-sensitive alignment-based feature extraction through two-phase frameworks. It consists of two parts. The first is the RRoI Learner, which learns the transition from Hroi to RRoI. The second is Rotated Position Sensitive RoI Align, which extracts rotated invariant features from RRoI for subsequent object classification and position regression.

Horizontal object detection has many limitations due to the arbitrary direction of the object. To solve this problem, Zhou et al. [92] presented the TS^4, a rotating object detector with a two-stage selection method and an anchor refinement module (ARM). The architecture of the TS^4Net is shown in Figure 19.

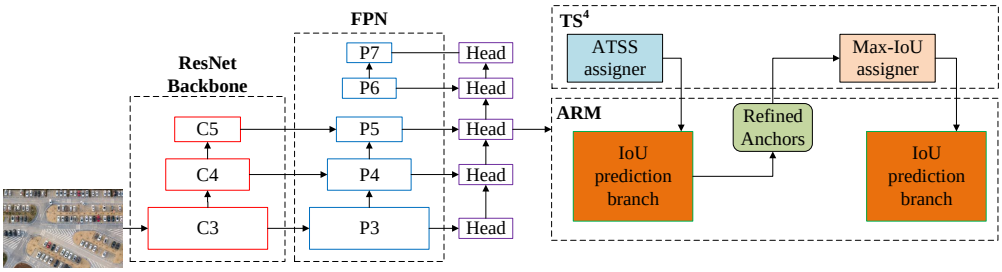

Figure 19. The architecture of the TS^4Net.

TS^4Net consists of a ResNet backbone, an FPN, an ARM cascade network with a rotated IoU prediction branch, and the two-stage sample selective strategy. In rotation detection, the angle causes a decrease in IoU value between the horizontal anchor and the ground truth. The rotation angle has a great effect on the IoU value, which makes the quality of the positive sample worse. To solve this problem, TS^4Net sets a single sample at each anchor point, greatly reducing calculations and increasing performance. Considering the ground truth contains the angle information of the object, TS^4Net utilizes the angle information between the sample and the ground truth to further select the high-quality candidate positive samples. The horizontal anchor is refined into a superior rotary anchor during the detection process by the first stage of ARM, and the rotary anchor is then adjusted into a more accurate prediction frame by the second stage.

The predicted results of car orientation are shown in Figure 20, where the red line indicates the head direction. Table 9 shows the quantity results with other state-of-the-art methods on UAV-ROD dataset. The results show that TS^4Net can achieve better accuracy in localization.

Object rotation is a major bottleneck in object detection for UAV images, and the realization of high-precision rotating object detection greatly expands the application scenarios of UAVs, especially the localization of targets in dense areas of a city in aerial images, such as vehicles in a parking lot, ships berthed in a harbor, and aircrafts in an airport.

Figure 20. Visualization results of the TS^4Net.

Table 9. Detection results on UAV-ROD dataset.

Method	Backbone	AP (%)	AP75 (%)	AP50 (%)
R-RetinaNet [102]	ResNet-50	71.46	85.88	97.68
R-Faster R-CNN	ResNet-50	75.79	86.38	98.07
TS^4Net	ResNet-50	75.74	86.9	97.82
TS^4Net	ResNet-101	76.57	88.17	98.1

4. Public Datasets for UAV Object Detection

Data-driven deep learning methods have been rapidly developed in recent years, providing powerful tools for object detection (images and videos) in the UAV remote sensing domain. To facilitate the research progress and performance evaluation of these tasks, many researchers have contributed various open-source and classic UAV-based remote sensing datasets. These datasets have a large scale and good generalization ability, which can reduce the characteristics of the dataset itself. In this section, we will introduce some of the most commonly used and influential datasets for UAV object detection.

Stanford Drone dataset [103]: This dataset was released by Stanford University's Computer Vision and Geometry Laboratory (CVGL) for studying human trajectory prediction and multi-object tracking in crowded scenes. The dataset contains eight different outdoor scenes, such as bookstores, cafes, campus squares, etc., each with multiple videos, totaling about 20,000 object trajectories. Each object's trajectory is annotated with a unique ID, containing 10 types of objects, more than 19,000 objects. Although this dataset only collects videos from a university campus, it can be applied to different application scenarios due to the diversity of scenes and the complexity of objects.

UAV123 dataset [104]: This dataset contains 123 video sequences captured from low-altitude UAVs, totaling more than 110,000 frames. These video sequences cover a variety of different scenes, such as urban areas, parks, beaches, and campuses, as well as different types of objects, such as pedestrians, bicycles, cars, and boats. Each video sequence has a corresponding bounding box annotation file that records the position and size of the object in each frame. In addition, each video sequence also has an attribute file that describes the features of the sequence.

Car Parking Lot dataset (CARPK) dataset [105]: The CARPK dataset is a dataset for vehicle detection and counting on UAV platforms. It was designated as a public dataset by National Taiwan University in 2017. Specifically, it is the largest and first parking dataset that UAV view has collected. The dataset was acquired by a UAV flying at an altitude of 40 m and includes images of nearly 90,000 vehicles taken from four different parking lots. The maximum number of vehicles in a single scene is 188. The label information of each vehicle is a horizontal bounding box.

UAV-ROD dataset [64]: In this dataset, 2806 aerial images were gathered using various platforms and sensors. Each image has items with sizes, shapes, and orientations and is roughly 4000 × 4000 pixels. These DOTA images are then annotated by aerial image interpretation experts using 15 common object categories. The fully annotated DOTA image contains 188,282 instances, each marked with an arbitrary (8d.o.F) quadrilateral.

Okutama-Action dataset [106]: The Okutama-Action Dataset is a video dataset for concurrent human action detection from a UAV perspective. It consists of 43 fully annotated sequences with a duration of one minute and contains 12 action categories. This dataset has many challenges that are lacking in current datasets, such as dynamic action transitions, significant scale and aspect ratio changes, sudden camera movements, and multi-label actors1. Therefore, this dataset is more challenging than existing datasets and will help advance the field and achieve real-world applications.

UAV Detection and Tracking (UAVDT) dataset [68]: The UAVDT dataset is a benchmark test for object detection and tracking on UAV platforms. It contains about 80,000 frames extracted from 10 h of videos involving three basic tasks, namely object detection (DET), single-object tracking (SOT), and multi-object tracking (MOT). The images in this dataset are taken by UAVs in various complex environments, with the main focus on vehicles. These images are manually annotated, including bounding boxes and some attributes that help analysis, such as vehicle category and occlusion. The UAVDT dataset consists of 100 video sequences that are filtered out from more than 10 h of videos taken at different locations in urban areas covering various common scenarios, such as square road toll stations, highway intersections, and T-junctions. The framerate of these videos is 30 frames per second (fps), and the resolution of the JPEG images is 1080 × 540 pixels.

DAC-SDC dataset [107]: The DAC-SDC dataset is a video dataset for object detection and classification on UAV platforms. It is provided by the System Design Contest (SDC) hosted by the IEEE/ACM Design Automation Conference (DAC) in 2019. The dataset is provided by Baidu and contains 95 categories and 150,000 images. These images are taken by UAVs in various complex scenarios, and each extracted frame contains 640 × 360 pixels.

Moving Object Recognition (MOR-UAV) dataset [108]: The Moving Object Recognition (MOR-UAV) dataset is a benchmark dataset for moving object recognition in UAV videos. The dataset contains 89,783 cars or heavy vehicle instances collected from 30 UAV videos, covering 10,948 frames. These videos are taken in different scenarios, such as occlusion, weather conditions, flying altitude changes, and multiple camera angles. The feature of this dataset is to use axis-aligned bounding boxes to annotate moving objects, which saves more computational resources than generating pixel-level estimates.

DroneVehicle dataset [108]: The DroneVehicle dataset is a benchmark dataset for vehicle detection and counting in UAV aerial images. The shooting environment covers from day to night, with real environment occlusion and scale changes. The dataset contains 15,532 pairs (31,064 images), half of which are RGB images and the other half are infrared images. It contains 441,642 annotated instances, divided into five categories: car, truck, bus, van, and cargo truck. Tilted bounding boxes are used to annotate vehicles to adapt to different angles and directions.

AU-AIR dataset [109]: The AU-AIR dataset is a large-scale object detection dataset of multimodal sensors captured by UAVs. It has 32,823 extracted frames from eight video sequences with different lighting and weather situations. The dataset contains eight types of objects, including people, cars, buses, vans, trucks, bicycles, motorcycles, and trailers. Each frame contains 1920 × 1080 pixels.

UVSD dataset [110]: The UVSD dataset is a large-scale benchmark dataset for vehicle detection, counting, and segmentation in UAV aerial images. The dataset consists of 5874 images with resolutions ranging from 960 × 540 to 5280 × 2970 pixels. The 98,600 vehicle instances in this dataset have accurate instance-level semantic annotations including three formats: pixel-level semantics, horizontal bounding boxes, and tilted bounding boxes. The dataset covers various complex scenarios, such as viewpoint changes, scale changes, occlusion dense distribution, and lighting changes, which result in great challenges for UAV vision tasks.

The above dataset is mainly used for urban environmental object detection. In recent years, natural disasters have had a profound impact on various regions around the world. The use of UAV geological object detection can make a significant contribution to accurate damage assessment. Here are some datasets for geological post-disaster object detection:

Maduo dataset [111]: Firstly, a UAV was used to capture the entire scene of the Maduo earthquake disaster area, and then photogrammetry technology was used to process individual images into a large digital orthophoto map (DOM) for further use. Finally, all cracks in the DOM are depicted through extensive manual annotations. Seismologists use these cracks to evaluate the stability of faults in the area and serve as samples for supervised deep learning methods. This dataset contains 382 DOMs, covering the entire area affected by the Maduo earthquake.

UNFSI dataset [112]: UNFSI is a dataset of road crack images captured by UAVs. Considering the effectiveness of the input image, the collected images were cleaned as follows: (1) they do not belong to common crack types. (2) They have severe interference crack features. (3) They have indistinguishable pixels. (4) They do not contain cracks. The total number of images is 5705, with a size of 4000 $\times$ 2250 pixels. To ensure the original size of crack feature values and facilitate the adaptation of target detection algorithms to actual UAV detection processes, all the original images were cropped into 640 $\times$ 640.

RescueNet dataset [113]: RescueNet is a high-resolution post-disaster dataset that includes detailed classification and semantic segmentation annotations. This dataset aims to facilitate comprehensive scene understanding after natural disasters. RescueNet comprises post-disaster images collected after Hurricane Michael, obtained using UAVs from multiple impacted regions.

In the past few years, with the development of precision agriculture, accurate counting and measurement can improve the accuracy of activities such as crop monitoring, precision fertilization, and yield prediction. However, manual counting in the field is both labor-intensive and time-consuming. UAVs equipped with RGB cameras can quickly and accurately facilitate this task. The datasets used for UAV agricultural object detection are as follows:

Zenodo dataset [114]: The Zenodo dataset was recorded using the UAV platform DJI Matrice 210 at a flight altitude of 3 m above the vineyard. Every flight records one side of the vineyard row. Grape berries are in the pea size (BBCH75) and string closure (BBCH79) stages. Two annotation types were used: MOTS to detect and track grape bunches, and COCO to detect berries. All the annotations were labeled using CVAT software. This dataset can be used for object detection and tracking on UAV RGB videos for early extraction of grape phenotypic traits.

Grape-Internet dataset [115]: The grape images in the Grape-Internet dataset are all from the network, including various red grape varieties, totaling 787 grape images. After data cleaning, all images are randomly cropped to different resolutions (from 514 $\times$ 460 pixels to 4160 $\times$ 3120 pixels) and manually annotated using LabelImg. The images have differences in shooting angle and shooting distance, as well as varying degrees of occlusion and overlap. The above differences greatly increase the difficulty of detection and bring greater challenges to the detection network.

MangoYOLO dataset [116]: The MangoYOLO dataset consists of 1730 images for mango object detection. The size of each image is 612 $\times$ 512 pixels, and each image contains at least one annotated mango. The annotation file contains the bounding box coordinates and mango variety names for each mango in XML format.

5. Future Directions

At present, the interest in UAV object detection algorithms is increasing, and the existing algorithms have achieved good detection results, but there are still issues that need to be addressed. The interference caused by complex backgrounds to the object detection task has been effectively suppressed, but the existing algorithms still have false alarms and missed detection problems in a dense environment or an environment with a large number of similar objects. The object detection algorithms based on the two-stage method have advantages in the accuracy of classification and regression. However, on the UAV platform, in order to meet the real-time requirements of image processing, the object detection algorithm needs to have a high processing speed, which places higher demands

on the parameter size and computational complexity of the network. Shadow appears when the object is disturbed by direct light from the light source. This increases the difficulty of object recognition and limits object detection. Due to differences in shape and position, objects may move differently or exhibit multiple postures based on their real-world rules. For example, pedestrians can walk, run, stand, or sit. At the same time, the height variation in UAVs can easily cause changes in the scale of the same object in the visual image, which can interfere with object detection.

Aiming at the above problems and the research in recent years, this paper undertakes the following discussions on the future research directions of UAV object detection based on deep learning.

(1) Rely on unsupervised or semi-supervised training. The existing multi-object detection datasets for UAVs are small, and the labeling cost is high. Unsupervised learning and semi-supervised deep learning network training methods can learn useful features and knowledge from unlabeled or a small amount of labeled data to achieve UAV object detection. In addition, pre-trained models from other fields or tasks can be used, such as image classification or object detection in natural scenes, to initialize or fine-tune UAV object detection models, thereby utilizing knowledge from the source domain or task to improve model performance.

(2) Data preprocessing algorithm. The effect of the deep learning method depends on the quality of the input data but cannot distinguish the data. The computational efficiency of the deep learning model can be improved by starting with data enhancement and reducing redundant features. Due to the limitations of UAV flight altitude and payload, problems such as object overlap, coverage, and displacement are inevitable. Generative adversarial networks and instance segmentation can effectively address these issues before object detection.

(3) Multimodal data. Multimodal data refers to data obtained from different sensors, such as visible light, infrared, and radar. Multimodal data can provide richer and more complete information, which helps to overcome the limitations and deficiencies of single-modal data. The application of multimodal data fusion is very wide, and there are some challenges in the data fusion process. Firstly, there are various problems with the data source: data quality issues, errors, formatting errors, incompleteness, etc. Secondly, there is also the problem of noise. Noise is not unique to multimodal data, but it creates new problems as each method can generate noise and potentially affect each other. There are also problems such as large data volume and inconsistent data. To address these issues, it is necessary to convert data from different sources into a unified format and resolution, thus promoting data fusion and processing.

(4) Introducing models with lower computational power requirements. Deep learning can achieve adaptive optimization by adjusting the learning rate, but, when the data or sample size is large, or when there are high requirements for convergence, a suitable algorithm can be chosen to optimize the structure and parameters of the net to improve the detection effect. As a data-driven approach, deep learning is not the best solution to solve a particular problem. A more targeted algorithm and reasonably allocated weights can be selected to accomplish the task flexibly and efficiently.

(5) Phenotype analysis. With the development of UAV technology in precision agriculture [114], high-precision real-time detection of crops is of great significance. Here, phenotype analysis refers to the use of UAVs, deep learning, and object detection to measure and evaluate the morphological, structural, physiological, and biochemical characteristics of crops, which can optimize planting management and improve crop quality [115]. Traditional yield estimation relies on manual experience, with low efficiency and accuracy, and cannot meet the fast and accurate prediction needs of large-scale planting enterprises. Considering the significant differences in shape and size among crop varieties, designing an end-to-end lightweight UAV object detection algorithm can improve the speed, accuracy, and reliability of phenotype analysis while reducing costs and errors.

6. Conclusions

UAV technology is currently in a period of rapid development, and UAV object detection has a broad research prospect. This paper reviews the development history of deep-learning-based UAV object detection methods from the two main technical approaches to UAV object detection: one-stage and two-stage methods. We sort out the research results in the field of deep-learning-based object detection regarding UAV considering the relevant perspectives in recent years. We summarize the advantages and shortcomings of the current methods in solving the problems in UAV object detection, such as the increase in small objects, complex background, object rotation, scale change, and category imbalance. Datasets in this field are also introduced. Through the above summarization and analysis, the subsequent development trend and further research direction are prospected. It is expected to provide valuable references. Looking forward, there is potential to design an increased number of end-to-end UAV object detection systems. This can be achieved by integrating the capabilities of UAV flight and mission planning. Such integration will not only enhance the intelligence of UAV object detection but also meet different application requirements and scenario conditions.

Author Contributions: Conceptualization, G.T. and J.N.; methodology, G.T.; validation, J.N., Y.G. and W.C.; writing—original draft preparation, G.T.; writing—review and editing, G.T., J.N., Y.Z., Y.G. and W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61873086) and the National Key R&D Program of China (2022YFB4703402).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets relevant to our paper are available online.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, J.; Chen, M.; Hou, S.; Wang, Y.; Luo, Q.; Wang, C. An Improved S2A-Net Algorithm for Ship Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2023**, *15*, 4559. [CrossRef]
2. Gao, T.; Niu, Q.; Zhang, J.; Chen, T.; Mei, S.; Jubair, A. Global to Local: A Scale-Aware Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615614. [CrossRef]
3. Gao, L.; Gao, H.; Wang, Y.; Liu, D.; Momanyi, B.M. Center-Ness and Repulsion: Constraints to Improve Remote Sensing Object Detection via RepPoints. *Remote Sens.* **2023**, *15*, 1479. [CrossRef]
4. Ampatzidis, Y.; Partel, V.; Meyering, B.; Albrecht, U. Citrus rootstock evaluation utilizing UAV-based remote sensing and artificial intelligence. *Comput. Electron. Agric.* **2019**, *164*, 104900. [CrossRef]
5. Li, L.; Huang, W.; Gu, I.Y.H.; Tian, Q. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.* **2004**, *13*, 1459–1472. [CrossRef] [PubMed]
6. Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.C.; Shi, P. Deep learning-based scene understanding for autonomous robots: A survey. *Intell. Robot.* **2023**, *3*, 374–401. [CrossRef]
7. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Kang, X.; Chanussot, J.; Benediktsson, J.A. Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep: Overview and Toolbox. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 60–88. [CrossRef]
8. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978. [CrossRef]
9. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. Uav-yolo, Small object detection on unmanned aerial vehicle perspective. *Sensors* **2020**, *20*, 2238. [CrossRef]
10. Yuanqiang, C.; Du, D.; Zhang, L.; Wen, L.; Wang, W.; Wu, Y.; Lyu, S. Guided Attention Network for Object Detection and Counting on Drones. In Proceedings of the 28th ACM International Conference on Multimedia, MM 2020, Virtual Event, 12–16 October 2020; pp. 709–717.
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]

13. Ni, J.; Shen, K.; Chen, Y.; Cao, W.; Yang, S.X. An Improved Deep Network-Based Scene Classification Method for Self-Driving Cars. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5001614. [CrossRef]
14. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
15. Wu, X.; Hong, D.; Ghamisi, P.; Li, W.; Tao, R. MsRi-CCF: Multi-scale and rotation-insensitive convolutional channel features for geospatial object detection. *Remote Sens.* **2018**, *10*, 1990. [CrossRef]
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016.
17. Ni, J.; Shen, K.; Chen, Y.; Yang, S.X. An Improved SSD-Like Deep Network-Based Object Detection Method for Indoor Scenes. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5006915. [CrossRef]
18. Luo, X.; Tian, X.; Zhang, H.; Hou, W.; Leng, G.; Xu, W.; Jia, H.; He, X.; Wang, M.; Zhang, J. Fast automatic vehicle detection in UAV images using convolutional neural networks. *Remote Sens.* **2020**, *12*, 1994. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861v1.
21. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767v1.
22. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9182–9192.
23. Wang, J.; Zhang, T.; Cheng, Y.; Al-Nabhan, N. Deep learning for object detection: A survey. *Comput. Syst. Sci. Eng.* **2021**, *38*, 165–182. [CrossRef]
24. Ni, J.; Chen, Y.; Chen, Y.; Zhu, J.; Ali, D.; Cao, W. A Survey on Theories and Applications for Self-Driving Cars Based on Deep Learning Methods. *Appl. Sci.* **2020**, *10*, 2749. [CrossRef]
25. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
26. Ahmad, H.M.; Rahimi, A. Deep learning methods for object detection in smart manufacturing: A survey. *J. Manuf. Syst.* **2022**, *64*, 181–196. [CrossRef]
27. Valavanis, K.P.; Vachtsevanos, G.J. *Handbook of Unmanned Aerial Vehicles*; Springer: Dordrecht, The Netherlands, 2015; pp. 1–3022.
28. Cazzato, D.; Cimarelli, C.; Sanchez-Lopez, J.L.; Voos, H.; Leo, M. A survey of computer vision methods for 2d object detection from unmanned aerial vehicles. *J. Imaging* **2020**, *6*, 78. [CrossRef] [PubMed]
29. Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 91–124. [CrossRef]
30. Storch, M.; De Lange, N.; Jarmer, T.; Waske, B. Detecting Historical Terrain Anomalies with UAV-LiDAR Data Using Spline-Approximation and Support Vector Machines. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3158–3173. [CrossRef]
31. Wu, S.; Wang, L.; Zeng, X.; Wang, F.; Liang, Z.; Ye, H. UAV-Mounted GPR for Object Detection Based on Cross-Correlation Background Subtraction Method. *Remote Sens.* **2022**, *14*, 5132. [CrossRef]
32. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. [CrossRef]
33. Mehmood, K.; Ali, A.; Jalil, A.; Khan, B.; Cheema, K.M.; Murad, M.; Milyani, A.H. Efficient online object tracking scheme for challenging scenarios. *Sensors* **2021**, *21*, 8481. [CrossRef] [PubMed]
34. Akshatha, K.R.; Karunakar, A.K.; Shenoy, S.; Dhareshwar, C.V.; Johnson, D.G. Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 77–89.
35. Li, X.; Diao, W.; Mao, Y.; Gao, P.; Mao, X.; Li, X.; Sun, X. OGMN: Occlusion-guided multi-task network for object detection in UAV images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 242–257. [CrossRef]
36. Liu, H.; Fan, K.; Ouyang, Q.; Li, N. Real-time small drones detection based on pruned yolov4. *Sensors* **2021**, *21*, 3374. [CrossRef]
37. Ye, T.; Qin, W.; Li, Y.; Wang, S.; Zhang, J.; Zhao, Z. Dense and Small Object Detection in UAV-Vision Based on a Global-Local Feature Enhanced Network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2515513. [CrossRef]
38. Liu, H.; Qiao, J.; Li, L.; Wang, L.; Chu, H.; Wang, Q. Parallel CNN Network Learning-Based Video Object Recognition for UAV Ground Detection. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 2701217. [CrossRef]
39. Zheng, J.; Fu, H.; Li, W.; Wu, W.; Yu, L.; Yuan, S.; Tao, W.Y.W.; Pang, T.K.; Kanniah, K.D. Growing status observation for oil palm trees using Unmanned Aerial Vehicle (UAV) images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 95–121. [CrossRef]
40. Xiao, J.; Zhang, S.; Dai, Y.; Jiang, Z.; Yi, B.; Xu, C. Multiclass Object Detection in UAV Images Based on Rotation Region Network. *IEEE J. Miniaturization Air Space Syst.* **2020**, *1*, 188–196. [CrossRef]
41. Li, Z.; Pang, C.; Dong, C.; Zeng, X. R-YOLOv5: A Lightweight Rotational Object Detection Algorithm for Real-Time Detection of Vehicles in Dense Scenes. *IEEE Access* **2023**, *11*, 61546–61559. [CrossRef]
42. Vinci, A.; Brigante, R.; Traini, C.; Farinelli, D. Geometrical Characterization of Hazelnut Trees in an Intensive Orchard by an Unmanned Aerial Vehicle (UAV) for Precision Agriculture Applications. *Remote Sens.* **2023**, *15*, 541. [CrossRef]

43. Zhu, H.; Yu, H.B.; Liang, J.H.; Li, H.Z. Improved algorithm of UAV search based on electric field model and simulation analysis. *Jilin Daxue Xuebao (Gongxueban)/J. Jilin Univ. (Eng. Technol. Ed.)* **2022**, *52*, 3029–3038.
44. Li, L.; Ren, J.; Wang, P.; Lyu, Z.; Sun, M.; Li, X.; Gao, W. Image enhancement method based on exposure fusion for UAV aerial photography. *Xibei Gongye Daxue Xuebao/J. Northwestern Polytech. Univ.* **2022**, *40*, 1327–1334. [CrossRef]
45. Cheng, B.; Ji, H.; Wang, Y. A new method for constructing roads map in forest area using UAV images. *J. Comput. Methods Sci. Eng.* **2023**, *23*, 573–587. [CrossRef]
46. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
47. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
48. Rejeb, A.; Abdollahi, A.; Rejeb, K.; Treiblmaier, H. Drones in agriculture: A review and bibliometric analysis. *Comput. Electron. Agric.* **2022**, *198*, 107017. [CrossRef]
49. Xu, X.; Dong, S.; Xu, T.; Ding, L.; Wang, J.; Jiang, P.; Song, L.; Li, J. FusionRCNN: LiDAR-Camera Fusion for Two-Stage 3D Object Detection. *Remote Sens.* **2023**, *15*, 1839. [CrossRef]
50. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. I511–I518.
51. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
53. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
54. Hossain, S.; Lee, D.J. Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with GPU-based embedded devices. *Sensors* **2019**, *19*, 3371. [CrossRef] [PubMed]
55. Lu, Y.; Guo, J.; Guo, S.; Fu, Q.; Xu, J. Study on Marine Fishery Law Enforcement Inspection System based on Improved YOLO V5 with UAV. In Proceedings of the 2022 IEEE International Conference on Mechatronics and Automation, ICMA 2022, Guilin, China, 7–10 August 2022; pp. 253–258.
56. Lalak, M.; Wierzbicki, D. Automated Detection of Atypical Aviation Obstacles from UAV Images Using a YOLO Algorithm. *Sensors* **2022**, *22*, 6611. [CrossRef] [PubMed]
57. Li, X.; Yang, J.; Li, Z.; Yang, F.; Chen, Y.; Ren, J.; Duan, Y. Building Damage Detection for Extreme Earthquake Disaster Area Location from Post-Event Uav Images Using Improved SSD. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpurs, Malaysia, 17–22 July 2022; pp. 2674–2677.
58. Sommer, L.W.; Schuchert, T.; Beyerer, J. Fast Deep Vehicle Detection in Aerial Images. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 311–319.
59. Liu, K.; Mattyus, G. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942.
60. Zhang, Y.; Wu, C.; Zhang, T.; Liu, Y.; Zheng, Y. Self-Attention Guidance and Multiscale Feature Fusion-Based UAV Image Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6004305. [CrossRef]
61. Audebert, N.; Le Saux, B.; Lefevre, S. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sens.* **2017**, *9*, 368. [CrossRef]
62. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
63. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Jie, Z.; Zhang, T.; Yang, J. Learning object-wise semantic representation for detection in remote sensing imagery. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; Volume 2019, pp. 1–8.
64. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
65. Zhang, Y.; Zhang, Y.; Fu, R.; Shi, Z.; Zhang, J.; Liu, D.; Du, J. Learning Nonlocal Quadrature Contrast for Detection and Recognition of Infrared Rotary-Wing UAV Targets in Complex Background. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 173–190. [CrossRef]
66. Uzkent, B.; Yeh, C.; Ermon, S. Efficient object detection in large images using deep reinforcement learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1813–1822.
67. Li, Z.; Liu, X.; Zhao, Y.; Liu, B.; Huang, Z.; Hong, R. A lightweight multi-scale aggregated model for detecting aerial images captured by UAVs. *J. Vis. Commun. Image Represent.* **2021**, *77*, 331–337. [CrossRef]
68. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 17–24 May 2018; Lecture Notes in Computer Science; Volume 11214, pp. 375–391.

69. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector with Spatial Context Analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1758–1770. [CrossRef]
70. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
71. Wang, J.; Shao, F.; He, X.; Lu, G. A novel method of small object detection in uav remote sensing images based on feature alignment of candidate regions. *Drones* **2022**, *6*, 292. [CrossRef]
72. Ševo, I.; Avramović, A. Convolutional Neural Network Based Automatic Object Detection on Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 740–744. [CrossRef]
73. Zeng, S.; Yang, W.; Jiao, Y.; Geng, L.; Chen, X. SCA-YOLO: A new small object detection model for UAV images. *Vis. Comput.* **2023**, 1–17. [CrossRef]
74. Qian, Y.; Wu, G.; Sun, H.; Li, W.; Xu, Y. Research on Small Object Detection in UAV Reconnaissance Images Based on Haar-Like Features and MobileNet-SSD Algorithm. In Proceedings of the 2021 International Conference on Cyber Security Intelligence and Analytics (CSIA2021), Shenyang, China, 8–13 July 2021; Advances in Intelligent Systems and Computing; Volume 1342, pp. 708–717.
75. Tian, G.; Liu, J.; Yang, W. A dual neural network for object detection in UAV images. *Neurocomputing* **2021**, *443*, 292–301. [CrossRef]
76. Chen, Z.; Wang, M.; Zhang, J. *Object Detection in UAV Images Based on Improved YOLOv5*; Springer: Cham, Switzerland, 2023; Volume 173, pp. 267–278.
77. Wu, C.; Liang, R.; He, S.; Wang, H. Real-Time Vehicle Detection Method Based on Aerial Image in Complex Background. In Proceedings of the China Conference on Command and Control, Beijing, China, 7–9 July 2022; Lecture Notes in Electrical Engineering; Volume 949, pp. 508–518.
78. Xi, Y.; Jia, W.; Miao, Q.; Liu, X.; Fan, X.; Li, H. FiFoNet: Fine-Grained Target Focusing Network for Object Detection in UAV Images. *Remote Sens.* **2022**, *14*, 3919. [CrossRef]
79. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered Object Detection in Aerial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; Volume 7, pp. 8310–8319.
80. Ju, S.; Zhang, X.; Mao, Z.; Du, H. A Target Detection Algorithm in the Aerial Image Guided by the Density Map of the Target Center Point. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Proceedings of the ICNC-FSKD 2021 17, Guiyang, China, 24–26 July 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 849–857.
81. Wang, T.; Qin, R.; Chen, Y.; Snoussi, H.; Choi, C. A reinforcement learning approach for UAV target searching and tracking. *Multimed. Tools Appl.* **2019**, *78*, 4347–4364. [CrossRef]
82. Yudin, D.; Skrynnik, A.; Krishtopik, A.; Belkin, I.; Panov, A. Object Detection with Deep Neural Networks for Reinforcement Learning in the Task of Autonomous Vehicles Path Planning at the Intersection. *Opt. Mem. Neural Netw. (Inf. Opt.)* **2019**, *28*, 283–295. [CrossRef]
83. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision ECCV 2016: 14th European Conference, Scottsdale, AZ, USA, 3–7 November 2016; Lecture Notes in Computer Science; Volume 9908, pp. 630–645.
84. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
85. Sun, W.; Dai, L.; Zhang, X.; Chang, P.; He, X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. *Appl. Intell.* **2022**, *52*, 8448–8463. [CrossRef]
86. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; Volume 76, pp. 8231–8240.
87. Shao, Z.; Cheng, G.; Ma, J.; Wang, Z.; Wang, J.; Li, D. Real-Time and Accurate UAV Pedestrian Detection for Social Distancing Monitoring in COVID-19 Pandemic. *IEEE Trans. Multimed.* **2022**, *24*, 2069–2083. [CrossRef] [PubMed]
88. Zhu, J.; Yang, G.; Feng, X.; Li, X.; Fang, H.; Zhang, J.; Bai, X.; Tao, M.; He, Y. Detecting Wheat Heads from UAV Low-Altitude Remote Sensing Images Using Deep Learning Based on Transformer. *Remote Sens.* **2022**, *14*, 5141. [CrossRef]
89. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
90. Xu, X.; Wang, L.; Shu, M.; Liang, X.; Ghafoor, A.Z.; Liu, Y.; Ma, Y.; Zhu, J. Detection and Counting of Maize Leaves Based on Two-Stage Deep Learning with UAV-Based RGB Image. *Remote Sens.* **2022**, *14*, 5388. [CrossRef]
91. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; Volume 2019, pp. 2844–2853.
92. Zhou, J.; Feng, K.; Li, W.; Han, J.; Pan, F. TS4Net: Two-stage sample selective strategy for rotating object detection. *Neurocomputing* **2022**, *501*, 753–764. [CrossRef]
93. Avola, D.; Cinque, L.; Diko, A.; Fagioli, A.; Foresti, G.L.; Mecca, A.; Pannone, D.; Piciarelli, C. MS-Faster R-CNN: Multi-stream backbone for improved Faster R-CNN object detection and aerial tracking from UAV images. *Remote Sens.* **2021**, *13*, 1670. [CrossRef]

94. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 23, pp. 936–944.
95. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Volume 137, pp. 764–773.
96. Liu, Z.; Gao, G.; Sun, L.; Fang, Z. Hrdnet: High-Resolution Detection Network for Small Objects. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 7896–7911.
97. Li, Q.; Sun, M.; Dong, L.; Gao, X.; Wang, Z.; Zhang, H. HCD-Mask: A multi-task model for small object detection and instance segmentation in high-resolution UAV images. In Proceedings of the 2022 IEEE International Conference on Industrial Technology (ICIT), Shanghai, China, 22–25 August 2022; Volume 2022, pp. 6344–6358.
98. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459. [CrossRef]
99. Zhang, R.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Object detection in UAV images via global density fused convolutional network. *Remote Sens.* **2020**, *12*, 1–17. [CrossRef]
100. Liu, S.; Zhang, L.; Lu, H.; He, Y. Center-Boundary Dual Attention for Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 267–278. [CrossRef]
101. Cores, D.; Brea, V.; Mucientes, M. Spatio-Temporal Object Detection from UAV On-Board Cameras. In Proceedings of the International Conference on Computer Analysis of Images and Patterns, Virtual Online, 8–13 August 2021; Lecture Notes in Computer Science; Volume 13053, pp. 143–152.
102. Wu, J.; Song, L.; Wang, T.; Zhang, Q.; Yuan, J. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In Proceedings of the 28th ACM international Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1570–1578.
103. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning social etiquette: Human trajectory understanding in crowded scenes. In Proceedings of the Computer Vision ECCV 2016: 14th European Conference, Scottsdale, AZ, USA, 8–16 October 2016; Lecture Notes in Computer Science; Volume 9912, pp. 549–565.
104. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for UAV tracking. In Proceedings of the Computer Vision ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 8–16 October 2016; Lecture Notes in Computer Science; Volume 9905, pp. 445–461.
105. Hsieh, M.R.; Lin, Y.L.; Hsu, W.H. Drone-Based Object Counting by Spatially Regularized Regional Proposal Network. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Volume 2017, pp. 4165–4173.
106. Barekatain, M.; Marti, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 2153–2160.
107. Xu, X.; Zhang, X.; Yu, B.; Hu, X.S.; Rowen, C.; Hu, J.; Shi, Y. DAC-SDC Low Power Object Detection Challenge for UAV Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 392–403. [CrossRef] [PubMed]
108. Mandal, M.; Kumar, L.K.; Vipparthi, S.K. MOR-UAV: A Benchmark Dataset and Baselines for Moving Object Recognition in UAV Videos. In Proceedings of the 28th ACM International Conference on Multimedia, Virtual Online, 12–16 October 2020; pp. 2626–2635.
109. Bozcan, I.; Kayacan, E. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 8504–8510.
110. Zhang, H.; Sun, M.; Li, Q.; Liu, L.; Liu, M.; Ji, Y. An empirical study of multi-scale object detection in high resolution UAV images. *Neurocomputing* **2021**, *421*, 173–182. [CrossRef]
111. Yu, D.; Ji, S.; Li, X.; Yuan, Z.; Shen, C. Earthquake Crack Detection From Aerial Images Using a Deformable Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4412012. [CrossRef]
112. He, X.; Tang, Z.; Deng, Y.; Zhou, G.; Wang, Y.; Li, L. UAV-based road crack object-detection algorithm. *Autom. Constr.* **2023**, *154*. [CrossRef]
113. Rahnemoonfar, M.; Chowdhury, T.; Murphy, R. RescueNet: A High Resolution UAV Semantic Segmentation Dataset for Natural Disaster Damage Assessment. *Sci. Data* **2023**, *10*, 913. [CrossRef]
114. Ariza, M.; Baja, H.; Velez, S.; Valente, J. Object detection and tracking on UAV RGB videos for early extraction of grape phenotypic traits. *Comput. Electron. Agric.* **2023**, *211*, 108051. [CrossRef]
115. Zhang, C.; Ding, H.; Shi, Q.; Wang, Y. Grape Cluster Real-Time Detection in Complex Natural Scenes Based on YOLOv5s Deep Learning Network. *Agriculture* **2022**, *12*, 1242. [CrossRef]
116. Pichhika, H.C.; Subudhi, P. Detection of Multi-varieties of On-tree Mangoes using MangoYOLO5. In Proceedings of the 2023 11th International Symposium on Electronic Systems Devices and Computing (ESDC), Sri City, India, 4–6 May 2023.