electronics

MDPI

Article

Target Object Detection from Unmanned Aerial Vehicle (UAV) Images Based on Improved YOLO Algorithm

Arunnehru Jawaharlalnehru [1,*], Thalapathiraj Sambandham [2], Vaijayanthi Sekar [1], Dhanasekar Ravikumar [3], Vijayaraja Loganathan [3], Raju Kannadasan [4], Arfat Ahmad Khan [5,*], Chitapong Wechtaisong [6,*], Mohd Anul Haq [7,*], Ahmed Alhussen [8,*] and Zamil S. Alzamil [7]

[1] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai 600026, India; vaijayanthisekar@gamil.com

[2] Department of Mathematics, SRM Institute of Science and Technology, Vadapalani, Chennai 600026, India; thalapathirajs@gmail.com

[3] Department of Electrical and Electronics Engineering, Sri Sairam Institute of Technology, Chennai 600044, India; dhanasekar.eee@sairamit.edu.in (D.R.); vijayaraja.eee@sairamit.edu.in (V.L.)

[4] Department of Electrical and Electronics Engineering, Sri Venkateswara College of Engineering, Sriperumbudur 602117, India; kannadasanr@svce.ac.in

[5] College of Computing, Khon Kaen University, Khon Kaen 40000, Thailand

[6] School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

[7] Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia; z.alzamil@mu.edu.sa

[8] Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia

* Correspondence: arunnehru.aucse@gmail.com (A.J.); arfatkhan@kku.ac.th (A.A.K.); chitapong@g.sut.ac.th (C.W.); m.anul@mu.edu.sa (M.A.H.); aa.alhussen@mu.edu.sa (A.A.)

check for updates

Citation: Jawaharlalnehru, A.; Sambandham, T.; Sekar, V.; Ravikumar, D.; Loganathan, V.; Kannadasan, R.; Khan, A.A.; Wechtaisong, C.; Haq, M.A.; Alhussen, A.; et al. Target Object Detection from Unmanned Aerial Vehicle (UAV) Images Based on Improved YOLO Algorithm. *Electronics* **2022**, *11*, 2343. https://doi.org/10.3390/electronics11152343

Academic Editor: Manohar Das

Received: 4 June 2022
Accepted: 20 June 2022
Published: 27 July 2022

Abstract: Aerial image-based target object detection has several glitches such as low accuracy in multi-scale target detection locations, slow detection, missed targets, and misprediction of targets. To solve this problem, this paper proposes an improved You Only Look Once (YOLO) algorithm from the viewpoint of model efficiency using target box dimension clustering, classification of the pre-trained network, multi-scale detection training, and changing the screening rules of the candidate box. This modified approach has the potential to be better adapted to the positioning task. The aerial image of the unmanned aerial vehicle (UAV) can be positioned to the target area in real-time, and the projection relation can convert the latitude and longitude of the UAV. The results proved to be more effective; notably, the average accuracy of the detection network in the aerial image of the target area detection tasks increased to 79.5%. The aerial images containing the target area are considered to experiment with the flight simulation to verify its network positioning accuracy rate and were found to be greater than 84%. This proposed model can be effectively used for real-time target detection for multi-scale targets with reduced misprediction rate due to its superior accuracy.

Keywords: image processing; convolutional neural network (CNN); YOLO; target detection; Darknet 19

1. Introduction

1.1. Background

The researchers have remarkable achievements in Computer vision due to its developments in deep learning algorithms, hardware requirements, and the obtainability of datasets. Notably, object detection is one of the most important research directions in computer vision and large numbers of target detection techniques have recently been developed. Because, the targets in a scene vary in size, it's more critical to detect and recognize them at varied scales. Small object detection and localization error are a hard and exciting challenge in the task of object identification that has attracted the interest of researchers, motivating them to enhance task performance in aerial object recognition and tracking to minimize human effort with increased efficiency. Although deep detection models were designed to handle challenges connected to broad object detection, they nonetheless contribute to the success of small object detection on a specific level. To improve the detection performance of targets with different sizes, a multi-scale target detection algorithm is essential using multi-scale detection training, with changing the screening rules of the candidate box.

1.2. Literature Review

Specifically, object detection is the most cited research task due to its broad application in several domains [1]. The unmanned aerial vehicles (UAV) have exposed their extraordinary impending for commercial, military, and civil-government applications in a broad range, notably infrastructure inspection, aerial photography, logistics, etc. [2]. The engagement of a UAV integrated with computer vision practices is absolutely beneficial for errands that necessitate distinctive conception and robust insight [2,3]. Deep learning with a convolutional neural network (CNN) in the extraction of high-level image features has several advantages [4]. Additionally, some of the researchers proposed a regional convolutional neural network (RCNN) with the VOC2012 data set [5], and the average accuracy of target detection mean Average Precision (mAP) increased notably. Also, the authors demonstrate that Fast RCNN and ultra-fast Faster RCNN [6,7] offer superior accuracy with relatively faster, and the frame rate can reach five frames/s. Redmonet al. [8] proposed YOLO that reaches the speed to detect video at 45 Frame/s. While increasing detection speed, YOLO sacrificed accuracy and found a new way to combine classification and localization for future research purposes.

Further, Liu et al. [9] and Redmon et al. [10] suggested YOLO and Single Shot Multi-Box Detector (SSD) map that the detection speed increased significantly and achieved satisfactory results for overall systems. For the target detection process in the VOC 2007 dataset, when the detection speed is 67 frame/s, the mAP reaches 76.8% in the field of target detection and achieves the best detection results. Target detection and aerial image positioning task have many similarities, and all need to target areas with fast and accurate positioning. From the perspective of human cognition and judgment position, the eye can see the scene; a human can quickly find and locate the target object, which is the target detection process that requires the computer to complete the task.

Similarly, during the flight, the pilot can roughly determine the aircraft's location according to the familiar target area on the ground, but aerial image positioning also needs to teach the computer to complete the task. In recent years, the target detection technology has matured, and the accuracy rate and detection speed have been significantly improved. Some of the recent works carried out by the researchers are illustrated in Table 1.

Table 1. Recent research works carried out by the researchers.

Ref. No	Methodologies	Inferences	Limitations
[11]	Single Shot Multi-Box Detector (SSD)	- Used for vehicle detection. - Deep learning approach was adapted. - Found to be faster than other compared models	- Sensitivity and precision scales were decreased.
[12]	RCNN and HOG	- Adapted to identify the human presence. - Deep learning schemes were used. - The overall performance showed superior compared with other conventional schemes.	- The proposed approach is not suitable for real-time scenarios.

Table 1. *Cont.*

Ref. No	Methodologies	Inferences	Limitations
[13]	GANet	- Used to detect the human and pose estimation. - Deep learning model was adapted. - It offered ideal accuracy in detection.	- Training descriptions are clutter-free, contextual, and inappropriate for real-time scenarios.
[14]	MobileNet and SSD	- Object detection was carried out using the proposed method. - Deep learning approach was considered. - It offered best better accuracy with the foreground and background attention model.	- This work failed to demonstrate multi-objects.
[15]	CNN	- Human action detection was carried out using the considered model. - Deep learning approach was adopted. - It detected the multiple actions of humans with good accuracy.	- Waving hand action was concentrated more than the other movements.
[16]	YOLO, SSD and RCNN	- Human detection and their counts were performed effectively. - Deep learning approach was used to study the performance. - It offered good accuracy in counting the humans.	- This work failed to focus on more crowd patterns with improved accuracy.
[17]	CNN	- This work demonstrated the recognition of human activity. - Deep learning approach was used. - This pipeline scheme improved the accuracy and speed of the detection.	- It failed to recognize the human action from different camera angles.
[18]	Human shape validation filter	- This work identified the birds using the proposed method. - Deep learning approach was considered. - The detection shows better accuracy than other methods.	- It showed some difficulties in detecting the smaller birds.
[19]	YOLOv4	- This model was adapted to detect the apple flower bud. - Deep learning approach was considered. - It helped to determine the heat requirement based on feasible detection.	- The classification performance showed average results.
[20]	Tiny YOLO	- This work provided optimized performance in both terms FPS and mAP. - Used the own dataset for training the data. - Applied to detect the pest using the proposed approach.	- This scheme offered less accuracy for low-resolution images.

Considering all the inferences and limitations, this work attempted to improve the detection accuracy using the YOLO network and helped to study the aerial image positioning. Using the YOLO network as the main body, the improvement measures are put forward through the target frame dimension clustering, classification network pre-training, multi-scale detection training, and changing the screening rules of the candidate frame to make it better adapt to the positioning task. The core problem of image localization is transformed into the target detection problem, and the target detection data set is made by selecting the flight test area and using the feature with apparent characteristics in the area as the target area. It is possible to locate the target area in the aerial image acquired in real-time. At the same time, two or more target areas of the judgment are adapted so that the relative positional relationship between the targets can greatly improve the positioning accuracy. However, the introduction of oblique projection to form an oblique image increases the difficulty of detection. Still, it can expand the range of aerial image storage so that a single aerial image appears as much as possible in the target area to improve positioning accuracy.

2. Methodology

2.1. Principle Feature Extraction Network Darknet 19

YOLO plus SSD network structure with a new design classification network, namely Darknet 19, acts as the basic model of the network. Before the YOLO model, most of the target detection framework was carried out using VGG 16 [21] to extract features, but it is more complex and offers computationally more intensive. The YOLO framework is similar to GoogleNet [22] in the network structure, but the calculation is lower than VGG 16, with a reduced accuracy rate compared with VGG 16. As a result, Redmon designed CNN with both complexity and accuracy to improve the detection performance of the network [8]. The final base model adapted is Darknet 19, containing 19 convolution layers and five top pooling layers. Like VGG 16, the network uses a large number of 3 × 3 convolution kernels; after each pool operation (Size 2 × 2, Step 2) and the number of channels also doubled. Based on the idea of Network In-network [23], the 1 × 1 convolution kernel is placed between the 3 × 3 convolution kernel to compress features and increase network depth. After each convolution layer, the bulk normalization operation is improved, and the Dropout operation is removed. The performance of the Darknet 19, AlexNet [24], and VGG 16 is shown in Table 2. Darknet 19 is in the Top 1% and Top 5% in the accuracy of 72.9% and 91.2%, respectively, which are higher than AlexNet and VGG 16. The time parameters (Central Processing Unit (CPU) and graphics processor (GPU) are slightly longer than that of AlexNet and VGG 16. The overall comparison shows that the performance of Darknet 19 is superior to others.

Table 2. Performance comparison of Darknet-19.

Model	Top—1/%	Top—5/%	GPU/ms	CPU/s
Alexnet	57	80.3	1.5	0.3
VGG-16	70.5	90	10.7	4.9
DarkNet-19	72.9	91.2	6.0	0.66

2.2. YOLO

YOLO structure and its improvement in the detection network use Darknet 19 as the primary model for feature extraction. The modification is done by replacing the final convolutional layer Darknet 19 network by adding three dimensions of 3 × 3. The number of channels with 1024 convolution layer; each convolution layer, after adding a size of 1 × 1 convolution layer, the output dimension that is required to detect the number. Compared with conventional YOLO, this modified structure removes the complete connection layer. The whole network is a convolution operation that retains the spatial information, i.e., each feature point and the resulting original map of each correspondence. And they are drawing on the ideas of anchors in the Faster RCNN [25] method for the target box dimension and clustering data set to determine the size and number of anchors. The category predictions in YOLO are no longer tied together with each cell but rather by using anchors to predict categories and coordinates simultaneously. Due to removing the fully connected layer, the model contains only Convolution and pooling layers, so the input size is flexible. When training, every few rounds can change the model input size, and therefore the model for different size image become robust. Every ten cycles, the model randomizes to obtain a new image dimension as input to proceed with the training. This rule forced the model to consider different input resolutions. The model is faster for small input sizes; YOLO can adjust the speed and accuracy when required. In the case of low resolution (288 Pixel × 288 Pixel) as illustrated in Table 3, YOLO's processing speed can reach up to 90 Frames/s in the case of accuracy and Fast RCNN flat results are obtained from [10]. In the case of high resolution, the mAP of YOLOv2 in the data set of VOC 2007 can achieve the best results.

Table 3. Performance comparison of object detection box with state-of-the-art results.

Detection Network	Fast R-CNN	Faster R-CNN VGG-16	Faster R-CNN Resnet	YOLO	SSD 300	SSD 500	YOLO 288 × 288	YOLO 544 × 544
mAP	70	73.2	76.4	63.4	74.3	76.8	69.0	78.6
FPS	0.5	7	5	45	46	19	91	40

2.3. Coordinate Transformation

The camera coordinate system (x_c, y_c, z_c) to the camera optical lens center 'S' as the origin, the z-axis is perpendicular to the imaging plane up to the positive direction, and x-axis and y-axis are parallel to the two sides of the imaging plane (Figure 1). The global coordinate system (X_g, Y_g, Z_g) uses the internationally adopted geocentric coordinate system to the Earth centroid as the coordinate origin of the WGS 84g coordinate system.

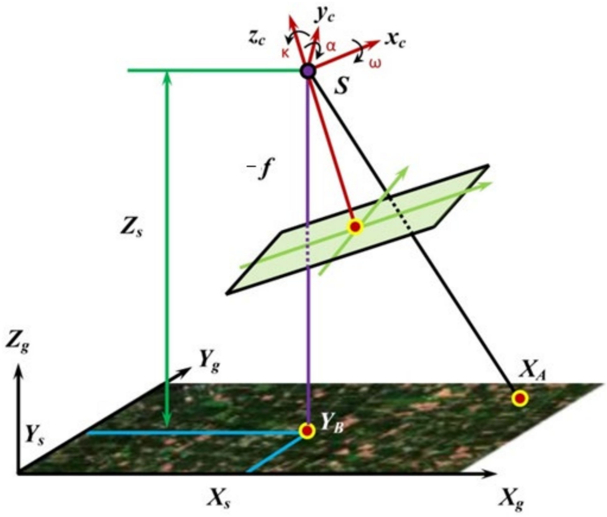


Figure 1. Imaging Model Diagram.

The external orientation elements of the imaging model comprise of 3 corner elements (α, ω, and κ) and 3 line elements (X_s, Y_s, Z_s), which are used to describe the spatial position coordinates of the camera's spatial posture and the Optical Center Point S. Global coordinates (X_g, Y_g, Z_g) and the camera coordinates (x_c, y_c, z_c) of the conversion is derived using the below formula [26]:

$$\begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} \begin{bmatrix} \cos\alpha\cos k - \sin\alpha\sin\omega\sin k & -\cos\alpha\sin k - \sin\alpha\sin\omega\sin k & -\sin\alpha\cos\omega \\ \cos\omega\sin k & \cos\omega\cos k & -\sin\omega \\ \sin\alpha\cos k + \cos\alpha\sin\omega\sin k & -\sin\alpha\sin k + \cos\alpha\sin\omega\cos k & \cos\alpha\cos\omega \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} + \begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} \tag{1}$$

where X_s and Y_s are the items to be solved and Z_s is the known item representing the flight altitude of the aircraft. Point A coordinates in the camera coordinate system are expressed as (X_A, Y_B, $-f$), where X_A, Y_A represents a point A location in aerial imagery; f represents a focal length which is a known term of the coordinates in the global coordinate system (X_A, Y_B, 0); wherein X_A and Y_B represent the coordinates of the target area as a known item and B states the midpoint of the plane.

The coordinates in the global coordinate system are X_S, Y_S, 0, according to the points S, A, B. The coordinate conversion relationship X can be derived between the positional as illustrated below:

$$[X_A Y_A Z_A] = M[x_A y_A - f] + [X_s Y_s Z_s] \tag{2}$$

In Equation (2), where the coordinates of the rotation matrix of 3 × 3 consist of three corner elements and M terms the aerial map. Further, the synthesis of both the equations can be calculated by X_S and Y_S, i.e., by calculating the projection relationship and the coordinate conversion. The coordinates of the onboard camera can obtain the coordinates of the center of the target area to obtain the coordinates of the centroid of the UAV.

3. Improved Method

Although YOLO has achieved the best detection results, it is not entirely suitable for image localization tasks. As shown in Figure 2, the basis of the YOLO network mainly improved by combining two different approaches as follows:

(1) The target frame of the self-made data set dimension clustering determines the anchor parameters of YOLOv2 are determined by clustering of VOC2007 and VOC2012datasets. The determined parameters of YOLOv2 are universal but not suitable for specific detection tasks; therefore, it is necessary to re-clustering operation in the self-made aerial image detection data set.

(2) Fine-tune the network using a different set of self-made data in classification network training. Like the YOLO, the first use of ImageNet dataset for pre-training, and the difference is the use of homemade resolution different image classification data set, can get a better fine-tuning effect.

(3) During the training process, every ten rounds change the input size of the model, and therefore, the model of different scales of the image becomes robust. The input data is self-made aerial image detection data set.

(4) Modify the filter rules of the candidate box to change the non-Maxima suppression (NMS) operation to the maximum value. The screening rule of the candidate frame in YOLO is NMS operation, but the maximum value operation can be carried out directly on the image positioning problem in this work to improve the detection effect.

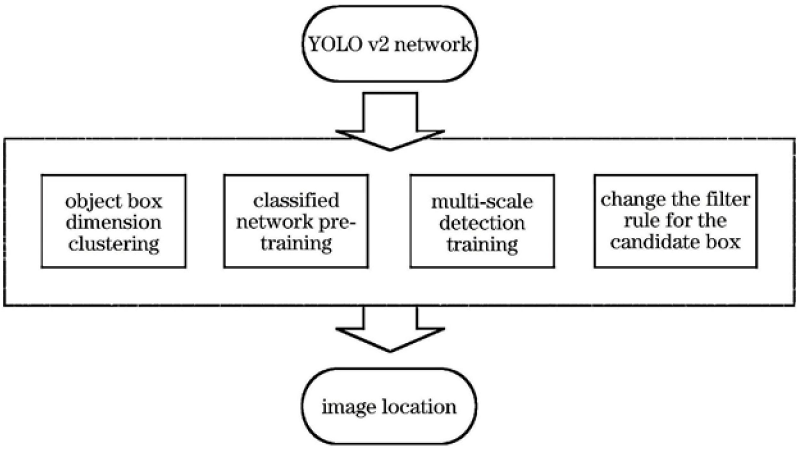
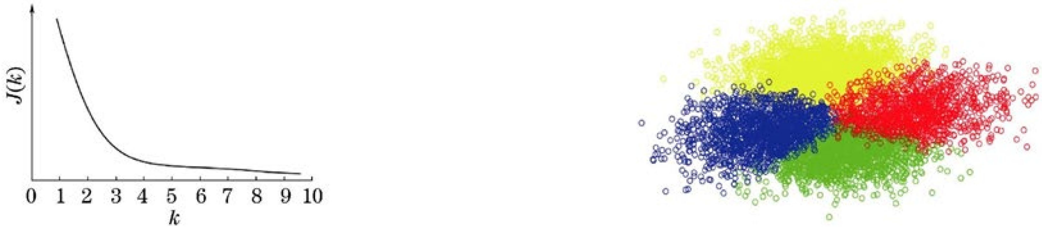

YOLO v2 network
object box dimension clustering
classified network pre-training
multi-scale detection training
change the filter rule for the candidate box
image location

Figure 2. Schematic of the improved method.

3.1. Target Box Dimension Clustering

The target box dimension clustering YOLO works on the idea of the Faster RCNN and introduces the anchors in a set of constant size and aspect ratio of the initial candidate box anchor. The accuracy of the target's detection concerning speed and the position of the target box, but the anchor dimensions are set manually. Redmon et al. [8] proposed a method of dimensionally clustering by k-Means for the data set manually labeled target box clustering and finds the statistical laws of the target box. YOLO has five anchors, and the number of clusters from VOC and COCO is also 5. The J function represents the sum of squares of the distance from each sample point to its cluster center, and the purpose of *k*-means is to adjust the *J* function to a minimum. When the number of clusters is increased to a particular value with the increase in the number of clusters, the objective function changes very little, and this inflection point can be considered the optimal number of clusters—using the *k*-means algorithm to cluster the width and height of the target frame in the data set, the objective function changes as shown in Figure 3a. When $k > 4$, the objective function changes very little; therefore, it takes four as the optimal number of clusters. When $k = 4$, the target block clustering distribution formed as shown in 3b, four different color regions corresponding to four different types of the target frame are observed, and the number of the anchor 'r' is 4; width and height dimensions are four color regions.



(**a**) (**b**)

Figure 3. (**a**) Objective Function Change Curve (**b**) Cluster of Objective Box.

The center point of the cluster corresponding to the target frame width and height in the configuration file changes the parameters anchors (11.68, 10.38), (18.11, 17.23), (9.59, 16.83) and (15.72, 20.75), corresponding to the green, red, blue, yellow cluster centers region.

3.2. Classification Network Pre-Training

Classification network pre-training is an essential part of target detection, and the ability and speed of the classification network are to extract features directly, which affect the target detection. A pre-trained version of the classification network on ImageNet [24] can be used as a framework for extracting features, but due to the limitations of the fully connected layer, the input data will be uniformly adjusted to a fixed size. The first step is to pre-train the Darknet 19 and this is done through an existing data set named ImageNet. Considering the pre-training's performance, processing power and efficiency are advisable to use low-resolution (224 × 224) pixel aerial images. The pre-training of these images on the ImageNet data set allows classification and fine-tuning of Darknet 19 to recognize aerial images exponentially.

The resolution in Darknet 19 can be changed to be higher 448 × 448 pixels and then let the weight of every single layer adapt to the resolution. The classification algorithm is switched to a detection algorithm better to adapt aerial image features and multi-scale detection tasks.

3.3. Multi-Scale Detection Training

Since YOLO contains only the Convolution and pooling layer, it is possible to change the size of the input image at any time. In the training process, the input size of the model is changed every ten rounds, and therefore the model has robustness to the images of different sizes. Since the down sampling factor of the model is 32, the size of the input image is a multiple of 32, and the size calculation can be performed using the below formula:

$$S = 32(7 + a) \tag{3}$$

where S is the size of the input image and a is a natural number randomly generated in the 0~12. This multi-scale training rule forced the model to adapt to different input resolutions. Compared with the fixed-resolution model, the multi-scale detection training is faster for low-resolution input image detection and more accurate for high-resolution input image detection.

3.4. Change the Candidate Box

Initially, the input Aerial image is divided into S × S grid blocks, and each block predicts the object present at the center of the block, where bounding boxes and their confidence scores are estimated. During the training process, each candidate box of YOLO will calculate their confidence as follows:

$$CS_{conf} = Pr(Object) \times R_{pred}^{truth} \times Pr\left(\frac{Class_i}{Object}\right) \tag{4}$$

The class specific confidence score, CS_{conf}, is defined as $Pr\left(O_{object}\right) *$ grid R_{pred}^{truth}, where $Pr\left(O_{object}\right)$ represents the probability that the block contains an object in the predicted bounding box and the target grid R_{pred}^{truth} helps to predict the rate of overlap between the bounding box and the ground truth. The block i the bounding box is also predicts the uncertain class probabilities, $\Pr\left(\frac{Class_i}{Object}\right)$, for class objects to determine which class the object in the bounding box belongs to. If the bounding box is appeared in the area where the object exists, then $Pr\left(O_{object}\right)$ takes 1; otherwise, takes 0.

$$Pr\left(O_{object}\right) * R_{pred}^{truth} = Pr(Class_i) * R_{pred}^{truth} \tag{5}$$

where $Pr\left(O_{object}\right)$ *and* $Pr(Class_i)$ represents each grid prediction category probability score. After the composite score of each candidate box is obtained, the threshold value of $T_{threshold}$ is set, and the candidate box with a low score is filtered out. Thus, the Non-maximum Suppression (NMS) operation is changed to the maximum value of the operation, i.e., taking the maximum value in several groups. It is more significant than the threshold value of the composite score uniquely by determining the position of the candidate frame and its prediction category. The maximum value of the operation in the aerial image of a class of target area can only detect a target box that can effectively avoid the target area and calculated as:

$$Pr(Class_i) = \ Pr\left(O_{objecti}\right) * R_{pred}^{truth} \geq T_{threshold}, \tag{6}$$

The false identification of a similar area caused by network detection improves target detection accuracy.

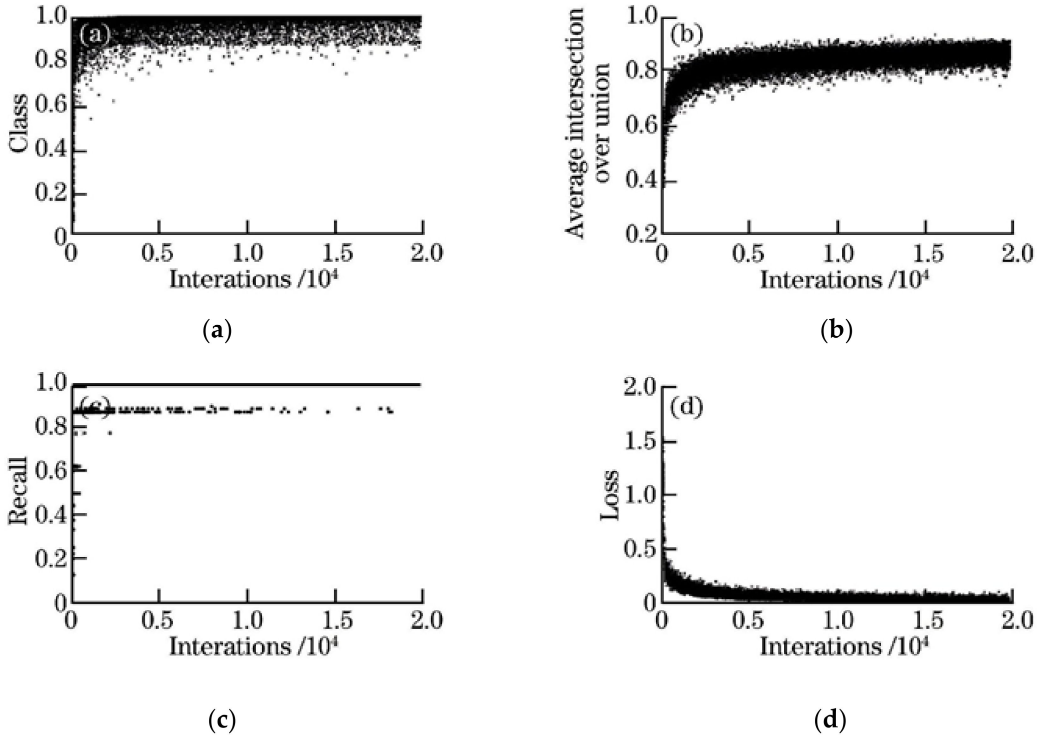
4. Results and Discussions

4.1. Experimental Data

The experimental data is based on the rectangular area centered in Changchun city, Jilin province. The source image was derived from Google Earth for April 2013, October 2015 and November 2016 in the city of Changchun in Jilin province Satellite Remote Sensing which lies in the middle portion of the Northeast China Plain. The orbital aerial image tile is captured from the altitude of 15 km above the sea level with an elevation of 200 m with different aerial angles. This data set is divided into two categories: (1) pre-training process that requires the use of classification data according to the resolution level which is divided into two groups, namely 224 pixel × 224 pixel and 448 pixel × 448 pixel; (2) detection network training that needs detection data according to the resolution and the study area was divided into 256 square areas of the same size with classification data set consists of 25 classes namely Agricultural land, Airport, Art gallery, Bus station, Church, Colleges, Film studio, Football, court, Highway, Hotel, Industry, Lake, Library, Mountains, Museum, Oil Mill, Open Mines, Park, Pond, Shopping mall, Tennis court, Theme park, Town hall, Train station and Zoo, which marked evenly distributed areas with obvious characteristics [27]. By rotating, adding noise, adjusting the tone, and other methods expand the number of samples [28]. Finally, the total number of samples of the classification data set is 530,440, with the ratio of the high and low resolution of about 3:1. The total number of samples of the test data is 38,200, with the number of samples of different resolutions being almost the same. Also, the ratio of the ortho image to the tilted image is 1:1 in order.

4.2. Configuration and Training Results

The experimental configuration is as follows: graphics for NvidiaGTX 1070; Intel CPU Core i7 6700; clocked at 3.40 GHz; memory is 32 GB; the operating system is Ubuntu 14.04. Further, the network parameter is as follows: learning rate is 0.0001; batch size is 64; steps were taken as 100, 20,000, 3,500,000; max_batches are 50,000; scales are 10, 0, 1, 0.1; momentu is 0.9; decay is 0.0005. As shown in Figure 4, the graph represents the number of iterations, average interactions over union; the range is about 0~20,000 times with good recall over 0.8 and reduced loss to 0.1. The ordinate of the scatter plot represents the four important parameters in the course of target detection network training: category accuracy rate, average overlap rate, recall rate, and loss value. With the increase in the number of iterations, category accuracy and recall rate gradually close to 1; the average overlap rate is stable at 0.883; the loss value dropped to about 0.1. From the convergence of the parameters, the network training results are ideal.



(**a**) (**b**) (**c**) (**d**)

Figure 4. Networking Training Parameters of the Convergence Scatter Plot. (**a**) Class. (**b**) Intersection. (**c**) Recall. (**d**) Loss.

4.3. Performance Comparison

4.3.1. Candidate Frame Generation Scheme

This work discusses the method in correspondence to the candidate frame generation schemes of Faster Region CNN and YOLO. As discussed earlier, clustering of the target block concerning the target region of the self-made data set and the optimal candidate block generation scheme of the self-made data set is obtained in Table 4. The Dimension clustering gives 0.83 average overlap rates, which obtains better performance when compared Faster Region based CNN and YOLO.

Table 4. Anchors and Overlap rate.

Candidate Box	Anchors Generated	Average Overlap Rate
Faster Region based-CNN	7	0.77
YOLO	5	0.79
Dimension Clustering	4	0.83

4.3.2. Classification Network Pre-Training Method Comparison

The classification network pre-training method is divided into three steps that compare each phase's impact on the network classification extraction feature capabilities. Three different stages of the pre-training network used to control other variables unchanged as a feature extractor and uses the same method to detect network training. Also, the comparison is carried out for the performance of three networks namely ImageNet with 448 × 448 pixels (blue line), ImageNet fine-tuned with low resolution of 288 × 288 pixels (red line) and ImageNet fine-tuned with low resolution of 224 × 224 pixels (yellow line). Fine to determine the effectiveness of the pre-training method. The effect of different pre-training methods is illustrated in Figure 5 through the multi-resolution fine-tuning. After the classification of the network in the detection task and found to be better for ImageNet fine-tuned with low resolution of 224 × 224 pixels (yellow line) because mAP value reached to 79.5 when compared to ImageNet with 448 × 448 pixels and ImageNet fine-tuned with low resolution of 288 × 288 pixels. The results show that the pre-training method of multi-resolution fine-tuning can significantly improve the feature extraction capability of classification networks.

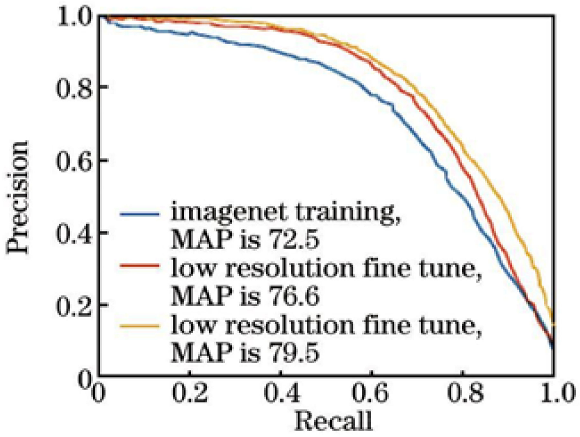


Figure 5. Comparison of different Pre-Training Methods.

4.3.3. Performance Measure between Different Networks

Performance measure of a multi-scale network is carried out by comparing a single-scale network through multi-scale network training that can show the detection of different scales of data sets. It offered strong adaptability through the detection of different scales of data sets and tested to obtain multi-scale network and single-scale network performance, as shown in Table 5. The input size of a single-scale network in Table 5 shows that the increase of the detection data set scale improved the detection effect of the two networks. Compared to a single-scale network with a multi-scale network, it is found that the smaller the scale of the detection data set, the faster the detection speeds. Also, the larger the scale of the detection data set, the higher the mAP value.

Table 5. Comparison of multi-scale and single-scale network performance.

Detection of Dataset/(Pixel × Pixel)	Multi-Scale Network		Single-Scale Network	
	Detection Time/s	mAP	Detection Time/s	mAP
224 × 224	0.01	71.1	0.013	70.3
320 × 320	0.012	74.8	0.014	74.2
416 × 416	0.015	77.5	0.015	77.8
512 × 512	0.018	79.4	0.016	78.4
608 × 608	0.029	80.9	0.018	78.9

4.3.4. Change Candidate Box Filtering Rules

To use the maximum value Operation MAX instead of NMS, change the candidate box filter rule is adapted. Different screening rules, namely MAX and NMS training network detection results are assessed and compared. It is found that the mAP value increased by 5 with NMS operation using the MAX rule screening network. So, the aerial image of a class of target area will only detect the most a target box, as far as possible to avoid the target area similar to the size of network detection. The comparison of different screening effects is shown in Figure 6.

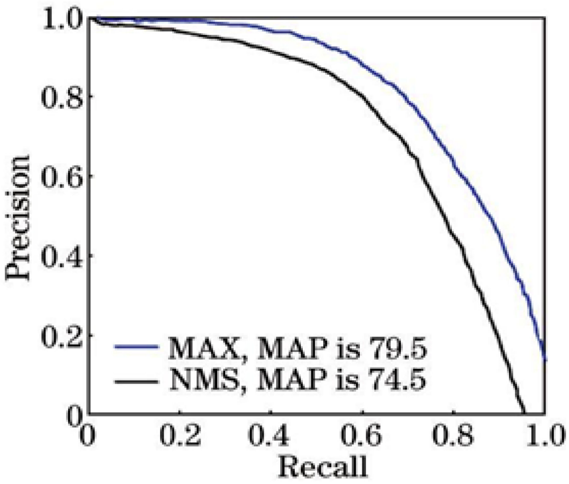


Figure 6. Comparisons of Different Screening Effects.

4.3.5. Determining Optimal Threshold and Detection Validation Set

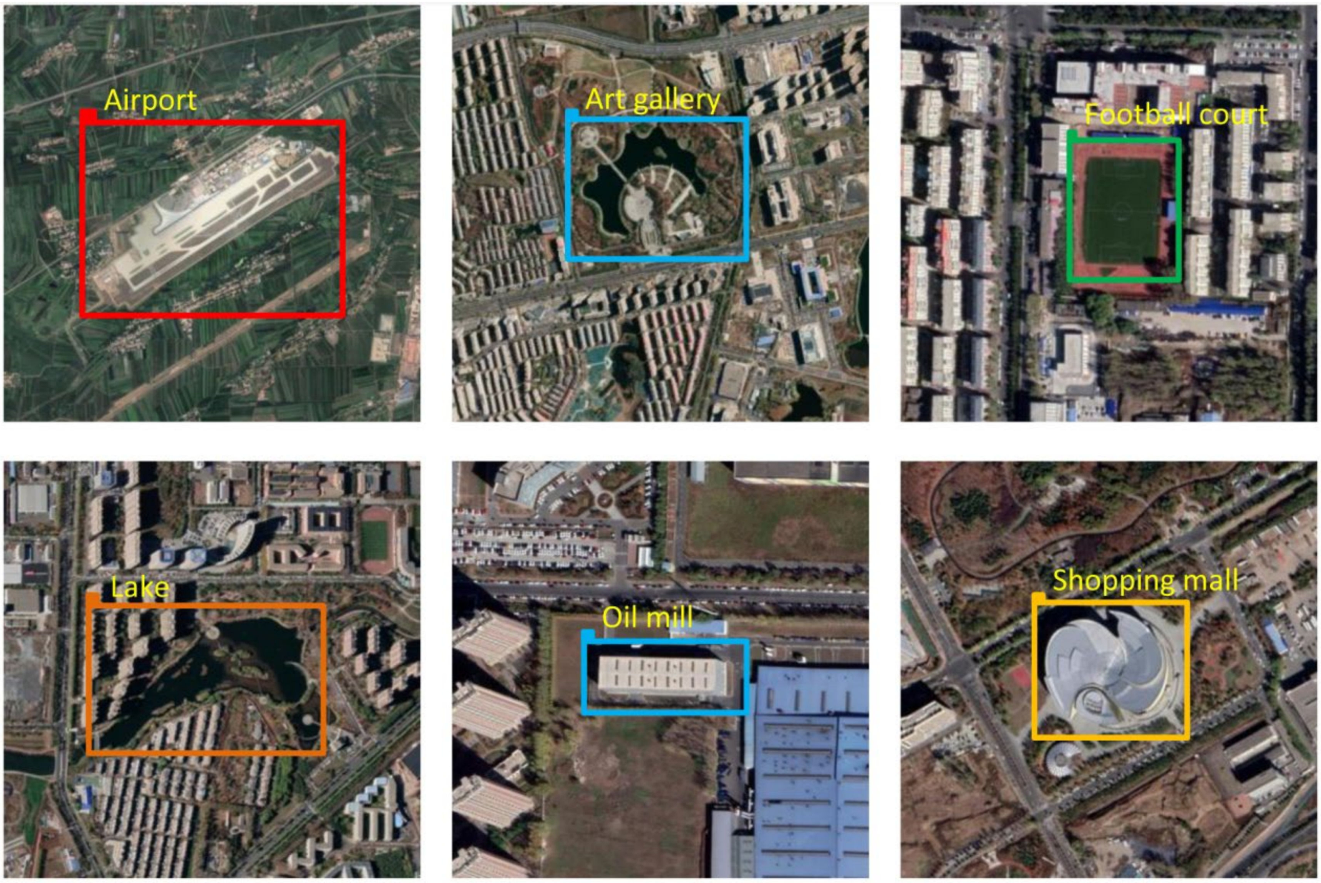
To determine the best threshold and test verification set in the detection process, setting the threshold to filter out the low-score candidate box, and different detection results are assessed. The improved method to train the detection network is compared with varying values of threshold and shown in Table 6.

Table 6. Comparison of the detection effects of different thresholds.

Threshold	Region Proposals/Image (Rps/Img)	Recall%	Precision%
0.001	7.75	100	12.07
0.005	3.13	86.78	50.27
0.025	1.56	69.91	80.45
0.45	1.17	50.29	93.73
0.8	0.12	12.03	100

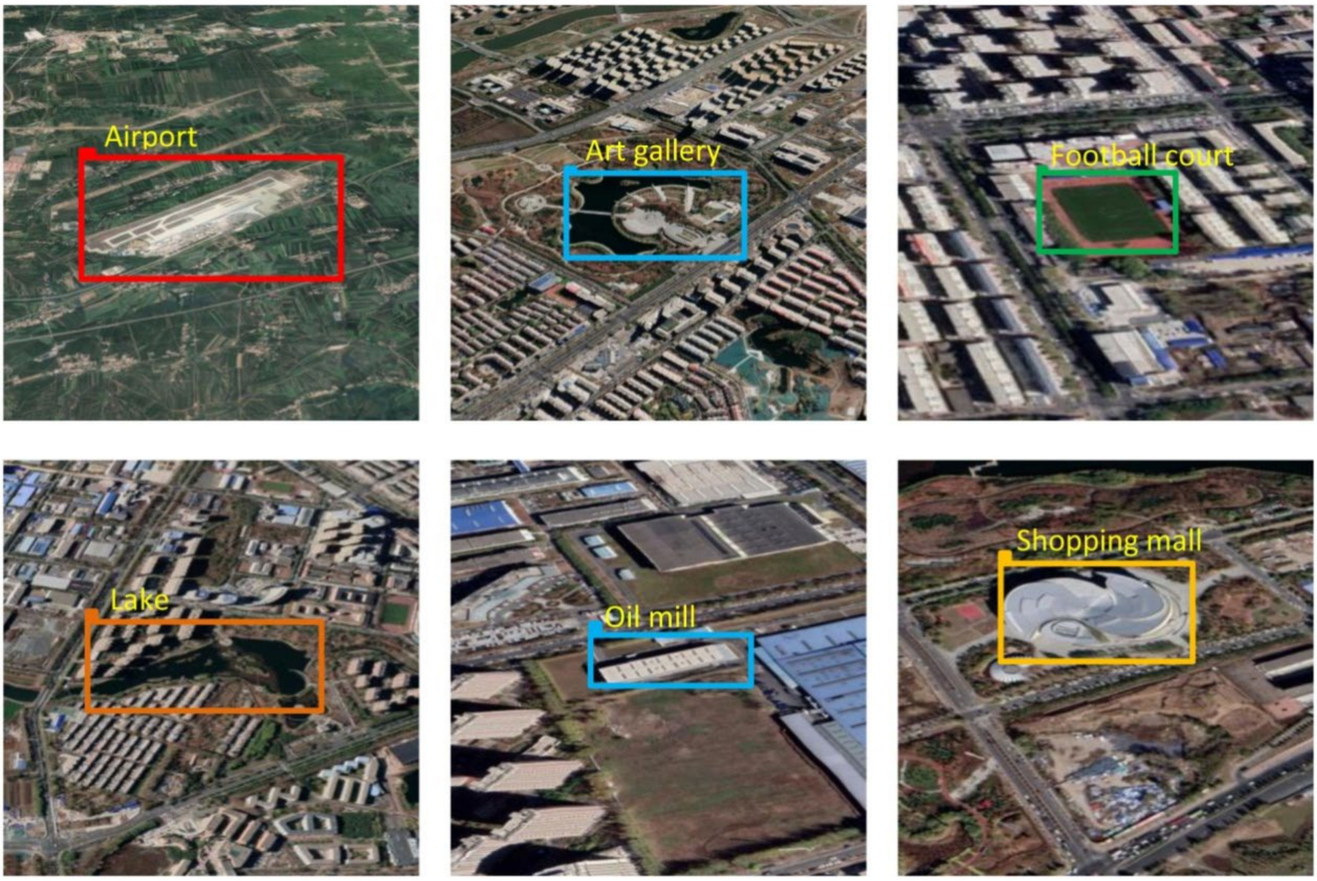
When using an improved method to complete the network training, the threshold is set to 0.025 with the validation set of samples to verify the effect of training and obtains results of precision as 80.45% and recall as 69.91%, which performs better when compared with other threshold values. Further, a part of the ortho results is shown in Figure 7a; ortho aerial image can accurately locate the position of the target area. Also, a part of the tilt image detection is shown in Figure 7b; and the results in Figure 7 correspond to this. For the same target area in aerial images, the different aerial angles will have a more significant deformation, and the network can be a good deformation of the target detected.

Furthermore, a part of the multi-target image detection results are shown in Figure 8; when there are two or more target areas in an aerial image, they can simultaneously detect multiple target areas [29–33]. As discussed earlier, the coordinate conversion and projection relationship can determine the latitude and longitude of the aircraft through the target frame mark aerial image position. When an aerial image contains multiple target frames, the need for comprehensive judgment takes place during multiple target frames are determined, the aircraft's latitude and longitude at the same time within the error range, and the output is positioned.



(**a**)

Figure 7. *Cont.*



(**b**)

Figure 7. (**a**) Detection Results of Orthographic Image (**b**) Detection Results of Sloping Image.

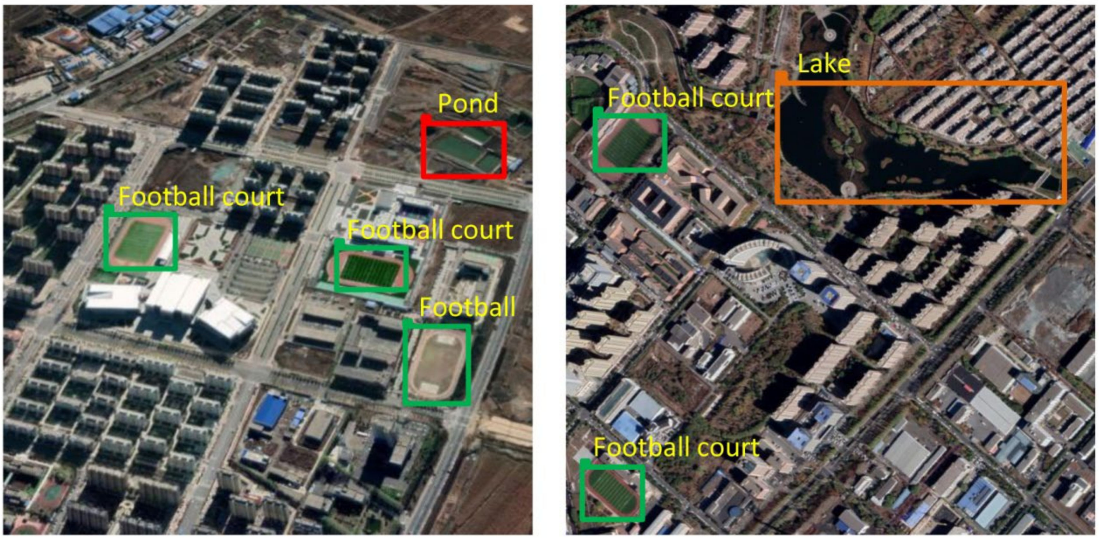


Figure 8. Detection Results of Multi-Target Image.

4.4. Simulation Verification

A simulation experiment was carried out under the Skyline environment of a high-resolution landscape to improve the method of training a good YOLO network; to determine whether the aerial image contains the target area and its location and category, Using projection and coordinate conversion relationship, real-time latitude and longitude of the UAV is assessed and then compared with the actual position of the UAV in Skyline. Select four different routes over the study area as the flight path of the UAV in the range of Changchun, on the four routes obtained.

Figure 9 represents a schematic diagram of Route 1. The camera takes four different tilt angles as a cycle for continuous aerial photography and takes four corresponding aerial images as a set sequence of camera pictures. Other routes in the image positioning effect are described in Table 7 within the scope of the study. The proportion of the target area image sequence contains an average of more than 40% in aerial images containing the target area and the positioning accuracy rate of more than 84%. The positioning accuracy rate can be increased when the aerial image includes a plurality of target areas.

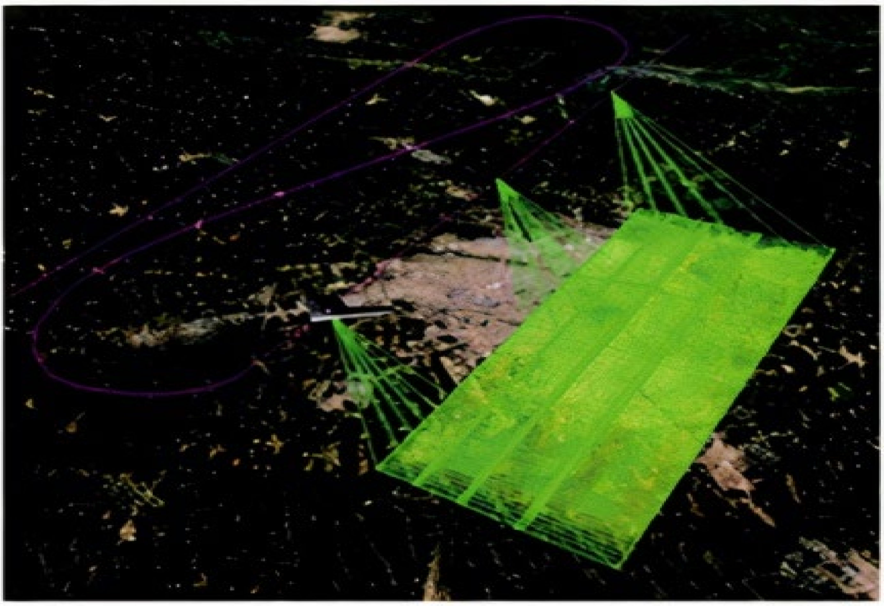


Figure 9. Diagram of Number 1 Route.

Table 7. Comparison of image positioning effects of different routes.

Track Number	Number of Sequences	Proportion of Object Area (%)	Number of a Correct Object Area	Number of a Wrong Object Area	Accuracy Rate of Location (%)
1	276	45	136	25	84.5
2	355	38	144	31	82.3
3	210	54	130	17	88.4
4	395	46	202	34	85.6

5. Conclusions

This work considered the YOLO network and improved the detection results utilizing the clustering of target frame dimensions, classification network pre-training, multi-scale detection training, and changing the screening rules of candidate frames. The proportion of the target area in the image sequence is about 40% on average. Further, the aerial images containing the target area offer a more than 84% positioning accuracy rate. This idea is to verify the feasibility of the image localization problem in target detection. But, there is also a small range of research data available from the existing study and the data sample production workload is insufficient for demonstration. Further, this research will be extended to adapt advanced methods to simplify data related to the hybrid network-based UAV image localization search.

Author Contributions: A.J., T.S., V.S., D.R. and A.A.K. proposed the idea under the supervision of the V.L., R.K., C.W., M.A.H., A.A. and Z.S.A. Moreover, V.L., R.K., C.W., M.A.H., A.A. and Z.S.A. contributed in the experimental simulations and designing, while A.J., T.S., V.S., D.R. and A.A.K. worked on the mathematical modeling. Also, A.J., T.S., V.S., D.R. and A.A.K. wrote the manuscript under the fruitful suggestions of V.L., R.K., C.W., M.A.H., A.A. and Z.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: Zamil S Alzamil would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-227.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Ramachandran, R.; Sangaiah, A.K. A review on object detection in unmanned aerial vehicle surveillance. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 215–228. [CrossRef]
2. 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] [PubMed]
3. Lo, L.-Y.; Yiu, C.H.; Tang, Y.; Yang, A.-S.; Li, B.; Wen, C.-Y. Dynamic Object Tracking on Autonomous UAV System for Surveillance Applications. *Sensors* **2021**, *21*, 7888. [CrossRef] [PubMed]
4. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
5. Girshick, R.; Donahue, J.; Darrel, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation C. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
6. 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. [CrossRef]
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster RCNN towards real time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified real time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
10. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
11. Kyrkou, C.; Plastiras, G.; Theocharides, T.; Venieris, S.I.; Bouganis, C.-S. DroNet: Efficient convolutional neural network detector for real-time UAV applications. In Proceedings of the 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, Germany, 19–23 March 2018. [CrossRef]
12. Yong, S.P.; Yeong, Y.C. Human object detection in forest with deep learning based on drone's vision. In Proceedings of the 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 13–14 June 2018; pp. 1–5.
13. Perera, A.G.; Al-Naji, A.; Law, Y.W.; Chahl, J. Human Detection and Motion Analysis from a Quadrotor UAV. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2018; Volume 405, p. 012003. [CrossRef]
14. Cai, Y.; Du, D.; Zhang, L.; Wen, L.; Wang, W.; Wu, Y.; Lyu, S. Guided attention network for object detection and counting on drones. *arXiv* **2019**, arXiv:1909.11307.
15. Mishra, B.; Garg, D.; Narang, P.; Mishra, V. Drone-surveillance for search and rescue in natural disaster. *Comput. Commun.* **2020**, *156*, 1–10. [CrossRef]
16. Gonzalez-Trejo, J.; Mercado-Ravell, D. Dense Crowds Detection and Surveillance with Drones using Density Maps. In Proceedings of the 2020 International Conference on Unmanned Aircraft Systems (ICUAS), Athens, Greece, 1–4 September 2020. [CrossRef]
17. Sien, J.P.T.; Lim, K.H.; Au, P.-I. Deep Learning in Gait Recognition for Drone Surveillance System. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019; Volume 495, p. 012031. [CrossRef]
18. Hong, S.-J.; Han, Y.; Kim, S.-Y.; Lee, A.-Y.; Kim, G. Application of Deep-Learning Methods to Bird Detection Using Unmanned Aerial Vehicle Imagery. *Sensors* **2019**, *19*, 1651. [CrossRef] [PubMed]
19. Yuan, W.; Choi, D. UAV-Based Heating Requirement Determination for Frost Management in Apple Orchard. *Remote Sens.* **2021**, *13*, 273. [CrossRef]
20. Chen, C.-J.; Huang, Y.-Y.; Li, Y.-S.; Chen, Y.-C.; Chang, C.-Y.; Huang, Y.-M. Identification of Fruit Tree Pests with Deep Learning on Embedded Drone to Achieve Accurate Pesticide Spraying. *IEEE Access* **2021**, *9*, 21986–21997. [CrossRef]
21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.

22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
23. Lin, M.; Chen, Q.; Yan, S.C. Network in network. *arXiv* **2013**, arXiv:1312.4400. [CrossRef]
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9. [CrossRef]
25. Zhang, S.; Zhao, H. Algorithm research of optimal cluster number and initial cluster center. *J. Appl. Res. Comput.* **2017**, *34*, 1617–1620.
26. Li, L. Terrain Reconstruction Based on Unmanned Aerial Vehicle Sequence Imaging and Its Application in Navigation. Ph.D Thesis, Changsha National University of Defense Technology, Changsha, China, 2009.
27. Zhang, L.; Chen, J.; Qiu, B. Region-of-Interest Coding Based on Saliency Detection and Directional Wavelet for Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 23–27. [CrossRef]
28. Liu, D.; Han, L.; Han, X. High spatial resolution remote sensing image classification base do depth learning Act. *Acta Opt. Sin.* **2016**, *36*, 0428001. [CrossRef]
29. Shu, C.; He, Y.; Sun, Q. Point Cloud Registration Based on Convolutional Neural Network. *Laser Optoelectron. Prog.* **2017**, *54*, 31001.
30. Arunnehru, J.; Geetha, M.K. A Quantitative Real-Time Analysis of Object Tracking Algorithm for Surveillance Applications. *Int. J. Emerg. Technol. Adv. Eng.* **2013**, *3*, 234–240.
31. Arunnehru, J.; Geetha, M.K. Motion Intensity Code for Action Recognition in Video Using PCA and SVM. In *Mining Intelligence and Knowledge Exploration*; Springer: Cham, Switzerland, 2013; Volume 8284, pp. 70–81.
32. Arunnehru, J.; Geetha, M.K. *Vision-Based Human Action Recognition in Surveillance Videos Using Motion Projection Profile Features*; Springer: Cham, Switzerland, 2015; Volume 9468, pp. 307–316.
33. Arunnehru, J.; Geetha, M.K. An efficient multi-view based activity recognition system for video surveillance using random forest. In *Smart Innovation, Systems and Technologies*; Springer: Cham, Switzerland, 2014; Volume 32, pp. 111–122.