aerospace

MDPI

Article

UAV Imagery for Automatic Multi-Element Recognition and Detection of Road Traffic Elements

Liang Huang [1,2], Mulan Qiu [1,*], Anze Xu [1], Yu Sun [1] and Juanjuan Zhu [1]

[1] Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China; kmhuangliang@kust.edu.cn (L.H.); 20202201076@stu.kust.edu.cn (A.X.); 20192201014@stu.kust.edu.cn (Y.S.); 20202201142@stu.kust.edu.cn (J.Z.)

[2] Surveying and Mapping Geo-Informatics Technology Research Center on Plateau Mountains of Yunnan Higher Education, Kunming 650093, China

* Correspondence: qiumulan@stu.kust.edu.cn

Abstract: Road traffic elements comprise an important part of roads and represent the main content involved in the construction of a basic traffic geographic information database, which is particularly important for the development of basic traffic geographic information. However, the following problems still exist for the extraction of traffic elements: insufficient data, complex scenarios, small targets, and incomplete element information. Therefore, a set of road traffic multielement remote sensing image datasets obtained by unmanned aerial vehicles (UAVs) is produced, and an improved YOLOv4 network algorithm combined with an attention mechanism is proposed to automatically recognize and detect multiple elements of road traffic in UAV imagery. First, the scale range of different objects in the datasets is counted, and then the size of the candidate box is obtained by the k-means clustering method. Second, mosaic data augmentation technology is used to increase the number of trained road traffic multielement datasets. Then, by integrating the efficient channel attention (ECA) mechanism into the two effective feature layers extracted from the YOLOv4 backbone network and the upsampling results, the network focuses on the feature information and then trains the datasets. At the same time, the complete intersection over union (*CIoU*) loss function is used to consider the geometric relationship between the object and the test object, to solve the overlapping problem of the juxtaposed dense test element anchor boxes, and to reduce the rate of missed detection. Finally, the mean average precision (mAP) is calculated to evaluate the experimental effect. The experimental results show that the mAP value of the proposed method is 90.45%, which is 15.80% better than the average accuracy of the original YOLOv4 network. The average detection accuracy of zebra crossings, bus stations, and roadside parking spaces is improved by 12.52%, 22.82%, and 12.09%, respectively. The comparison experiments and ablation experiments proved that the proposed method can realize the automatic recognition and detection of multiple elements of road traffic, and provide a new solution for constructing a basic traffic geographic information database.

Keywords: road traffic elements; channel attention mechanisms; UAV imagery; YOLOv4

Citation: Huang, L.; Qiu, M.; Xu, A.; Sun, Y.; Zhu, J. UAV Imagery for Automatic Multi-Element Recognition and Detection of Road Traffic Elements. *Aerospace* **2022**, *9*, 198. https://doi.org/10.3390/aerospace9040198

Academic Editor: Mostafa Hassanalian

Received: 9 February 2022
Accepted: 4 April 2022
Published: 6 April 2022

1. Introduction

Information on road traffic elements, including road centerlines, road intersections, zebra crossings, bus stations, roadside parking spaces, etc. are an important part of roads. The accurate recognition and detection of road traffic elements provide an essential decision-making basis for automatic driving, improving intelligent transportation systems, promoting smart cities, and updating basic traffic geographic information databases [1]. For the automatic recognition and detection of road traffic elements, the recent research of most scholars has been based on the detection and recognition of roadside traffic signage of a single element [2–4]. Inevitably, this approach has many shortcomings, such as the small amount of information acquired, the single element, and the large interval distance. This approach cannot provide a good solution for updating the basic traffic geographic

information database. Due to the limited shooting range, traditional vehicle-mounted cameras can obtain only a small portion of the road traffic element information. This is not conducive to the acquisition of large-area traffic element information; alternatively, unmanned aerial vehicles (UAV) images have the advantages of convenient acquisition and high resolution, providing favorable conditions for the acquisition of large-area traffic element information. Therefore, the automatic recognition and detection of multiple road traffic elements are studied through UAV remote sensing images in this paper to improve the efficiency and reduce labor costs for updating the basic traffic geographic information database.

Many studies have been carried out on target detection and recognition. With the development of deep learning, target detection methods have started changing from classical machine learning methods to deep learning methods, representing a new paradigm of machine learning. Target detection has been widely used in face detection [5], automatic driving [6], text detection [7,8], and other fields. Traditional target detection methods are based on color or shape features for target extraction. For example, Li et al. [3] proposed the method of detecting traffic signs through color and shape features; however, this method had a poor recognition effect and insufficient overall detection accuracy. Zhao et al. [4] used the Hough transform and shape analysis to detect and recognize road traffic signs; however, this method had an insufficient recognition rate and poor recognition effect. Berkaya et al. [9] used a shape algorithm and color threshold technology to detect and recognize circular traffic signs; however, this method only realized the recognition and detection of circular traffic signs, and its application scope was limited. Shi et al. [10] used a split-space Hough transformation method to achieve road boundary detection, and this method was suitable for boundary detection algorithms for straight and curved roads in general scenes. It is difficult to detect road boundaries in complex environments. He et al. [11] used shape information to detect triangular traffic signs; this method was only suitable for the detection of clear objects and did not detect the presence of fragments or occlusions in natural scenes. Creusen et al. [12] proposed an extended algorithm for traffic sign detection using information from multiple color channels. Most of these traditional methods use the special color and shape of traffic signs for feature extraction and rely on classifiers for classification. These methods generally suffer from slow detection speeds and insufficient detection accuracies, making it difficult to achieve the desired goal.

With the development of deep learning [13], increasing numbers of scholars are using deep learning for target detection. Target detection based on deep learning can be divided into two types: one-stage detection represented by a single-shot detector (SSD) [14], with a "You Only Look Once" (YOLO) algorithm [15–19], and two-stage detection represented by a region-based convolutional neural network (R-CNN) [20], Fast R-CNN [21], Faster R-CNN [22], etc. For example, a small traffic sign detection algorithm based on an improved SSD was proposed by Shan et al. [23], which achieved high accuracy in the test set but was not very applicable to the detection of other road traffic elements. Chen et al. [24] proposed an improved Mask R-CNN method to achieve road traffic sign recognition; however, this the method had a single recognition element and little information. Lodhi et al. [25] proposed a convolutional neural network (CNN)-based traffic sign recognition system. The authors integrated multilayer convolutional features and multilayer contextual information through a CNN framework for feature extraction. Guo et al. [26] used Faster R-CNN to implement a systematic approach for end-to-end traffic sign recognition. The method had good performance in small target detection and classification. Jin et al. [27] proposed an improved solution to the problem of insufficient average detection accuracy and missed detections during target detection in real road scenes. The authors improved the detection accuracy of road targets with the YOLOv3 improvement algorithm.

The problems and solutions proposed by the above scholars are useful for updating the basic traffic geographic information database on transportation. However, most scholars perform target detection based on a single element, which cannot satisfy the practical application needs for the detection of multielement road traffic. Therefore, a YOLOv4 [15]

network improvement algorithm combining the attention mechanism of efficient channel attention (ECA) [28] is proposed in this paper to achieve the automatic recognition and detection of multielement road traffic in UAV images. First, this paper manually labels a set of road traffic multielement datasets on UAV images captured by roLabelImg [29] (downloadable from https://github.com/cgvict/roLabelImg, accessed 23 June 2020). Second, the optimal candidate box size of the target object is obtained by k-means clustering analysis. Then, by integrating the ECA mechanism into the YOLOv4 backbone network, dataset training is conducted to detect the accuracy of multiple road traffic elements. At the same time, the complete intersection over union (*CIoU*) loss function [30] is introduced to reduce the error detection rate of juxtaposed dense elements side by side and greatly improve the detection accuracy.

In order to recognize automatic multiple elements and detect road traffic, it is possible to provide a service for updating the basic traffic geographic information database. The main contributions of this paper are as follows:

(1) In response to the problems related to few road traffic multielement datasets, single elements, and lack of road information, a set of UAV image road traffic multielement datasets are produced in this paper.

(2) Aiming to solve the problem of the insufficient detection accuracy of road elements and the difficult identification of juxtaposed dense elements, the YOLOv4 algorithm integrating the ECA mechanism is proposed.

(3) The comparative experiment and ablation experiment prove the superiority of this method in detecting multiple elements of road traffic and provide a new solution for updating the basic traffic geographic information database.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 details the proposed method, followed by the experiments and results in Section 4. The discussion is presented in Section 5. Finally, our conclusion is outlined in Section 6.

2. Related Work

In recent years, UAVs have shown a wide range of advantages in the field of transportation. In particular, they play an important role in road traffic monitoring, navigation, road damage detection, vehicle tracking for identification, road maintenance, and other traffic components [31–42]. With the advantages of fast data collection, high image quality, minimal cost, light weight, and great adaptability, UAVs can be used in road traffic inspection to greatly improve efficiency and reduce maintenance and manpower costs.

Research in the field of transport drones has focused on the problem of cruise route planning for UAVs, road vehicle detection, and the extraction of road information. The following scholars have addressed the problem of UAV cruise-route planning. Liu et al. [31] proposed a multi-objective optimization model for UAV cruise path planning. Additionally, an improved algorithm was designed to solve the UAV cruise path planning problem. Cheng et al. [32] proposed an algorithm for optimizing and modifying the optimal paths for UAVs. The authors also developed a multibase, i.e., a rechargeable and refillable UAV road patrol task allocation model to solve the problem of poor endurance associated with UAVs. Other academics have contributed to the traffic control field by completing real-time monitoring of road traffic information through UAVs. Elloumi et al. [33] proposed a road traffic detection system based on several UAVs. The authors monitored the traffic on urban roads with several UAVs in real time and sent the information to a traffic processing center for traffic control. Yang et al. [34] proposed an artificial intelligence-based solution to implement multi-object detection for intelligent road monitoring. The method provides a good solution for future road monitoring and control of intelligent transportation by combining UAVs, wireless communication, and Internet of Things technologies. Wang et al. [35] proposed a method for handling the loss of contact between a UAV and its operator based on probabilistic model detection. The method enables the UAV to perform surveillance tasks in dangerous environments. Huang et al. [36] proposed a distributed navigation scheme. This scheme achieved road traffic condition detection in different modes by means

of real-time UAV detection. Liu et al. [37] proposed a real-time UAV rerouting model and a decomposition-based multi-objective optimization algorithm. The model took the dynamic requirements of traffic monitoring into account to achieve dynamic route planning for UAV cruising, making it more suitable for real-life traffic monitoring. UAV technology has the advantages of low cost, high flexibility, and good quality of collected image data. The growing number of UAV applications in the field of transport is reflected based on the increasing amount of road image information collected with UAVs. Pan et al. [38] detected asphalt pavement deterioration through drone imagery to provide decision support for road maintenance practices. The paper proposes that a combination of machine learning algorithms, such as support vector machines, artificial neural networks, and random forests, can be used to differentiate between normal and damaged pavements for pavement damage identification. Saad et al. [39] used UAV images to identify ruts and potholes in road surfaces. The authors identified the ruts and potholes extracted from UAV images through site survey and planning, data acquisition, data processing and results, and data analysis to achieve road condition detection. Roberts et al. [40] proposed a method for generating 3D pavement modeling using UAV images. These models were used to monitor and analyze the pavement condition and to automate the detection of pavement deterioration. Wang et al. [41] proposed a UAV-based target tracking and recognition system. This system implemented the functions of target tracking, target recognition and detection, and image processing. Liu et al. [42] processed UAV images through a target detection network with multiscale feature fusion, improving the ability to detect small targets while reducing resource consumption makes the network lighter.

The above paragraph describes the main research on UAVs in the field of transportation. Similar to the abovementioned scholars, this paper also collects information through UAVs. Most scholars currently collect road image information through UAVs primarily to research road damage. However, this paper mainly collects information on road traffic elements, including road centerlines, road intersections, zebra crossings, bus stations, roadside parking spaces, and other similar information. A review of a large amount of literature shows that there is little research on the automatic identification and detection of road traffic elements. However, the extraction of road traffic element information is of great significance for updating the basic traffic geographic information database. Determining how to extract road traffic elements in a low-cost and high-efficiency way is particularly important. Therefore, this paper proposes a deep learning method by fusing the multielement images of road traffic obtained by UAVs. It can achieve the effect of automatic identification and detection with high efficiency, low cost, and high accuracy. The proposed method can provide technical support for updating the basic traffic geographic information database.

3. Research Method

YOLOv4 is an algorithm that combines a number of previous research techniques, combined with innovation. YOLOv4 enables efficient target detection tasks while using only a single GPU. In the YOLOv4 network, the training process can be optimized to improve accuracy. Better performance can also be achieved by sacrificing a little amount of inference time. The YOLOv4 network achieves the perfect balance of speed and accuracy in target detection tasks.

YOLOv4 has the advantages of fast detection and high speed. Therefore, a YOLOv4 network incorporating the ECA mechanism [28] is proposed. First, the k-means [43] clustering method is used to calculate the matching candidate box size in the datasets. Second, the mosaic data augmentation method is used to increase the number of training samples for multiple elements of road traffic on the training datasets. Then, the ECA module is fused into the YOLOv4 network for data training. Finally, the detection results are obtained, and the accuracy is evaluated. Figure 1 displays a flow chart of the proposed method.

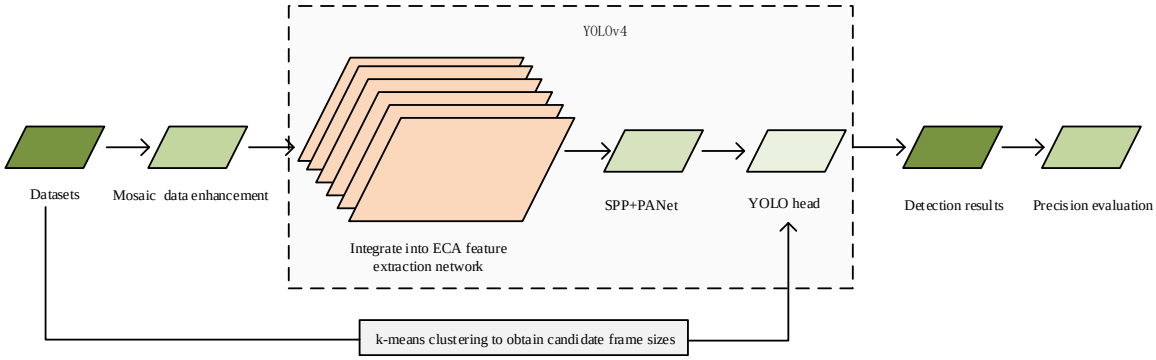


Figure 1. Flow chart of proposed method.

3.1. Overall Framework

In this paper, the YOLOv4 network model is selected as the base algorithm model. Bochkovskiy et al. [15] proposed that the YOLOv4 network is a one-stage target detection network. The YOLOv4 network is primarily composed of components related to the CSP-Darknet53 [15], spatial pyramid pooling (SPP) [44], feature pyramid networks (FPNs) [45], and path aggregation network (PAN) [46]. Among these components, the CSPDarknet53 structure consists of 5 content security policy (CSP) [47] modules, which are made to act as downsampling modules using a convolutional kernel with a step size of 2 and a size of 3×3 in front of each CSP module. Thus, when the input feature image is 416 pixels $\times$ 416 pixels in size, the image is downsampled after 5 CSP modules to obtain a feature map with a size of 13×13. CSPDarknet53 reduces the computational consumption and memory costs while also enhancing the learning capability of the CNNs and ensuring computational accuracy. The SPP structure is mainly used to solve the problem of the nonuniform size of the input image. The SPP structure directly pools the feature maps of any size to obtain a fixed number of features. FPN+PAN draws on the approach of PANet [46] by adding a feature pyramid to the tail of the FPN structure. This includes the two PAN structures to enable bottom-up communication of strong localization features, enabling easier reception of bottom-level information at the top of the hierarchy, and top-down communication of enhanced semantic features in combination with the FPN structure layer. The combination of these two components enables feature aggregation from different backbone layers and between detection layers, thus improving the feature extraction capability in the backbone network.

The YOLOv4 method of the fused ECA mechanism is proposed in this paper, which adds the ECA mechanism to the two effective feature layers extracted from the backbone network and to the result after upsampling, as shown in the YOLOv4 model with ECA in Figure 2.

3.2. Datasets and Scale Statistics

3.2.1. Introduction to the Datasets

At present, there is a lack of sufficient datasets for multiple road traffic elements, and most of the existing publicly available datasets are roadside traffic signage datasets or road traffic datasets. To meet the demand for updating the basic traffic geographic information database and to solve the problem of an insufficient number of datasets for road traffic elements, a set of datasets for multiple road traffic elements is produced in this paper, including zebra crossings, roadside parking spaces, and bus stations, as shown in the sample datasets in Figure 3. The UAV images were captured by the Hava MEGA-V8 and DJI FC6310. Harwar MEGA-V8 is equipped with a five-tilt camera, supporting the BeiDou, global positioning system (GPS), GLONASS, and seven real-time kinetic (RTK) Samsung frequencies. The horizontal positioning accuracy reaches $\pm$2 cm, and the vertical positioning error reaches $\pm$5 cm. This equipment is characterized by high efficiency, long endurance, and high-precision map formation. The DJI FC6310 UAV has 6 vision sensors, a

main camera, 2 sets of infrared sensors, 1 set of ultrasonic sensors, a GPS/GLONASS dual-mode satellite positioning system, an inertial measurement unit (IMU), and compass dual redundant sensors. This equipment can help the drone acquire real-time images and depth and positioning information while flying, as well as build a 3D map around the vehicle and determine its position. The image size of remote sensing is 7146 pixels × 5364 pixels and 5472 pixels × 3648 pixels, respectively. The spatial resolutions are 0.05 m and 0.1 m, respectively. A total of 16,872 images were taken, duplicate areas and areas without road traffic elements were removed from the images, and 1128 of these images were finally selected manually as the original dataset. The road traffic elements were manually marked with the image labeling software roLabelImg, which is used to mark rotated rectangular boxes or square rectangular boxes. The function used in this article involves the marking of positive rectangular boxes. There are many elements that constitute road traffic information, including road centerlines, road intersections, zebra crossings, bus stations, and roadside parking spaces. The main purpose of this paper is to achieve the automated construction of a basic traffic geographic information database. The research results of the automatic identification and detection of road traffic multi-elements are relatively few. Therefore, the representative traffic road elements are selected as research objects. Likewise, in this paper, zebra crossings, roadside parking spaces, and bus stations are selected as research objects. More types of automated detection and recognition will also be added in subsequent research. Among the research objects in this work, zebra crossings, roadside parking spaces, and bus stations are named zebra_crossings, parking_spaces, and bus_stations, respectively. The training data account for 90% of all data, and the rest are test data.

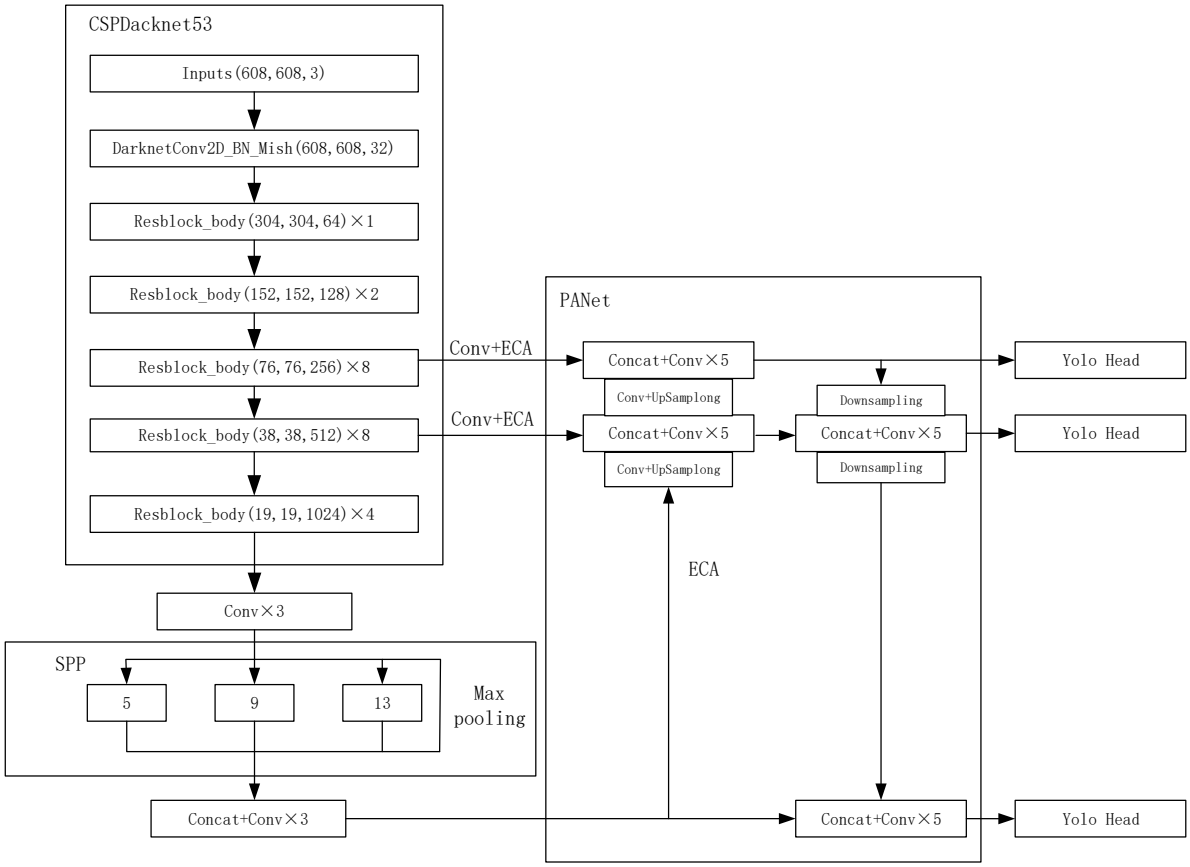


Figure 2. The YOLOV4 model with ECA.

Figure 3. Sample datasets.

3.2.2. Clustering of the Anchor Box

The anchor box sizes of the original YOLOv4 network were obtained from the visual object class (VOC) datasets [48]. The detection was performed for scales of 19 × 19, 38 × 38, and 76 × 76. The preset candidate boxes were 12, 16, 19, 36, 40, 28, 36, 75, 76, 55, 72, 146, 142, 110, 192, 243, 459, and 401, whose scale sizes are not applicable to the multielement datasets of road traffic captured by UAV images in this paper. Therefore, to apply the target scale range of road traffic multielement datasets, the k-means clustering method was used to conduct scale statistics on 1128 UAV road traffic multielement remote sensing images. First, the scale of road traffic elements was defined as 9 clusters, and the cluster centers of each cluster were randomly selected in each cluster. Then, each data point was associated to the nearest cluster center, and the center point of each of these 9 clusters was found as the new cluster center. Thus, the cluster centers were iterated until the points owned by these 9 clusters no longer change. Finally, the size of the target candidate box was set based on the clustering result. The k-means results clustering are shown in Table 1. These results show that the effect after clustering is in line with the target scale of the datasets proposed in this paper.

Table 1. The results of k-means.

Serial No.	1	2	3	4	5	6	7	8	9
x	11	13	17	25	28	57	60	96	117
y	22	14	33	77	20	119	45	99	59

3.3. Data Augmentation

Mosaic data augmentation is used to enhance the training datasets in the YOLOv4 network. The mosaic data augmentation approach starts by randomly extracting four images containing the anchor frames of the detectors from the road traffic multielement datasets; stitching the images into a new image by randomly scaling, cropping, and arranging them; obtaining the anchor boxes corresponding to this resulting image; and then passing this processed image into the YOLOv4 network for learning. The data from the four images can be calculated as one image for the batch normalization calculation [30]. As shown in the workflow of mosaic data augmentation in Figure 4, such mosaic data augmentation enriches the datasets with background and small sample information of the detection object. Moreover, mosaic data augmentation training does not require high computational performance, even when using only the central processing unit (CPU).

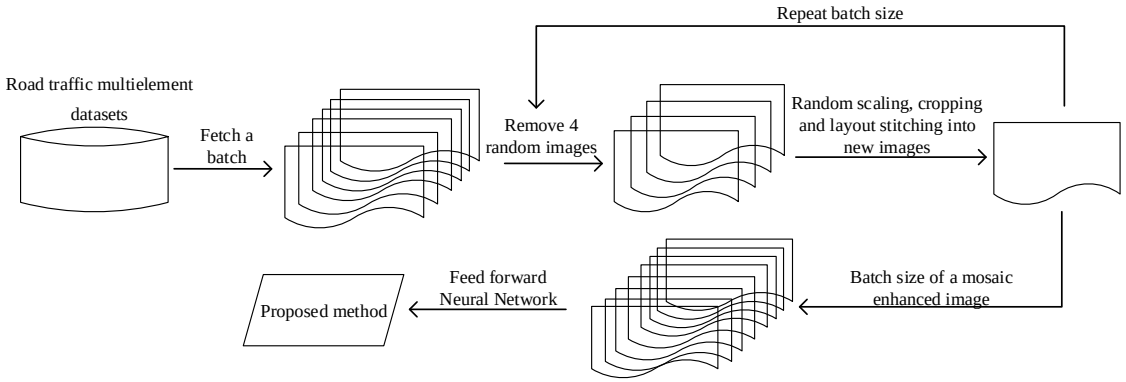


Figure 4. Workflow of mosaic data augmentation.

3.4. Efficient Channel Attention

In deep learning, the attention mechanism is a commonly used method and skill. There are many ways to realize the attention mechanism, but its core is to make the network focus on feature information. Attention mechanisms can be divided into channel attention mechanisms, spatial attention mechanisms, and a combination of the two. The mechanism used in this paper is the ECA mechanism. A local cross-channel interaction strategy without dimensionality reduction was implemented by one-dimensional convolution as well as an adaptive selection of the one-dimensional convolutional kernel size. With this method, the coverage of local cross-channel interactions can be guaranteed, which allows the network to gain performance improvements while reducing the complexity of the model.

ECANet [28] is an implementation of the channel attention mechanism. ECANet can be considered an improved version of SENet [49]. The squeeze-and-excitation (SE) [49] attention mechanism first carries out channel compression on the input feature map; but this dimension reduction method is not conducive to learning the dependency between channels. Therefore, the ECA avoids dimensional reduction, uses one-dimensional convolution to efficiently realize local cross-channel interaction, extracts the dependency between channels, and improves the performance of the YOLOv4 network. This likewise improves the identification accuracy of the road traffic elements in UAV images. The specific steps of ECA's attention mechanism are as follows:

(1) Create a feature map for the global averaging pooling operation.

(2) Carry out a one-dimensional convolution operation with a convolution kernel size equal to k and obtain the weight ω of each channel through the sigmoid activation function. The calculation formula of ω is:

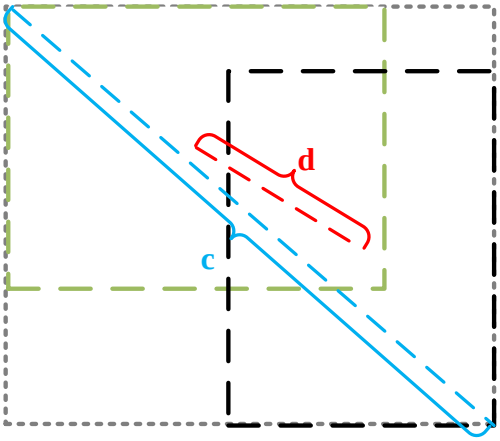
$$\omega = \sigma(C1D_k(y)) \tag{1}$$

where $C1D$ stands for one-dimensional convolution and k stands for the related parameter information between the corresponding y and k fields.

(3) The weights are multiplied by the corresponding elements of the original input feature map to obtain the final output feature image.

3.5. CIoU_Loss

Road traffic elements in UAV images, such as roadside parking spaces, have juxtaposed dense elements. The intersection over union (IoU) loss function is not a good solution to this problem; therefore, the *CIoU* loss function is used to solve this problem. *CIoU* [30] improves the function regression accuracy and convergence speed by considering the distance between the detection frame and target box, overlapping area, aspect ratio, and other aspects, as shown in Figure 5.



c: The length of the diagonal of the minimum external rectangle.
d: The distance between the centre points of the real box and the predicted box.

Figure 5. Diagram of *CIoU*.

CIoU, whose penalty items are publicly announced as:

$$R_{CIoU} = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{2}$$

where v is the similarity of the metric aspect ratio and α is the weighting function, respectively, defined as:

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{3}$$

$$\alpha = \frac{v}{1 - IoU} \tag{4}$$

Thus, the *CIoU* loss function can be expressed as:

$$CIoU\_Loss = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{5}$$

where c denotes the diagonal distance between the prediction box b and the smallest outer rectangle of the real box b^{gt}, and d denotes the distance between the centroid of the real box and the prediction box. IoU is the area intersection ratio of the prediction box and the real box. $\rho^2\left(b, b^{gt}\right)$ denotes the Euclidean distance between the prediction box and the centroid of the real box.

4. Experimental Results and Analysis

4.1. Experimental Environment

The computer configuration used was an i7-9700k CPU running Windows 10 with a GTX1070Ti GPU and 8 GB of video memory. The experimental training platform was Pycharm. The training weight decay coefficient was set to 0.0005, the initial learning rate was set to 0.001, the confidence level was set to 0.5, and the IoU threshold was set to 0.5. A total of 100 epochs were trained, with 4000 iterations. The datasets were divided into a training set and a validation set in a 9:1 ratio, and a typical road traffic element was randomly selected as the test set.

4.2. Evaluation Indicators

In the experiment, the mean average precision (*mAP*) was calculated as the quantitative evaluation index of the model to measure the accuracy of the model detection. The *mAP* is defined as:

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{6}$$

where N represents the number of all categories in the test set, i is the ith category, and AP_i is the average precision (AP) of the ith category, which is defined as:

$$AP = \int_0^1 p(r)d_r \tag{7}$$

where p is the precision; r is the recall; and p is a function with r as an argument, which is equal to taking the area under the curve. The *recall* and *precision* are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where TP represents the positive samples detected correctly, characterizing the number of road traffic elements detected correctly; FP represents the negative samples detected incorrectly, characterizing the number of targets the were incorrectly detected as classes other than road traffic element classes; and FN represents the positive samples detected incorrectly, characterizing the number of other classes detected incorrectly as road traffic element classes.

4.3. Comparison Experiments

In this paper, the effectiveness of the proposed method is verified for both classical and state-of-the-art algorithmic networks for target detection. The SSD, RetinaNet, Faster R-CNN, YOLOv3, YOLOv4, and YOLOv5 networks were used in comparison experiments to train the road traffic multielement datasets, and their AP, precision, recall, and mAP values were calculated and compared. As shown in Table 2, the recognition accuracy of road traffic elements under different network models was counted separately. The rise points in Table 2 are the mAP calculated by comparing each network with the proposed methods in this paper.

Table 2. Detection results of different models.

Network Model	Transport Elements	AP	Precision	Recall	mAP	Rise Points
Faster R-CNN	zebra crossings	64.26	59.04	71.43	56.89	33.56
	bus stations	71.46	73.33	70.97		
	roadside parking spaces	34.96	31.51	48.75		
Retinanet	zebra crossings	70.01	87.82	61.16	57.27	33.18
	bus stations	67.25	86.36	61.29		
	roadside parking spaces	34.54	74.28	24.38		
SSD	zebra crossings	52.92	76.84	32.59	53.94	36.51
	bus stations	75.09	**100**	54.84		
	roadside parking spaces	33.81	73.37	14.74		
YOLOv3	zebra crossings	84.25	87.38	80.36	81.52	8.93
	bus stations	83.81	88.89	77.42		
	roadside parking spaces	76.49	76.13	75.86		
YOLOv4	zebra crossings	81.82	89.95	79.20	74.65	15.80
	bus stations	76.77	90.48	61.29		
	roadside parking spaces	65.35	70.24	70.99		
YOLOv5	zebra crossings	93.61	**91.82**	93.50	86.84	3.61
	bus stations	73.82	68.51	69.42		
	roadside parking spaces	**93.10**	**91.41**	**90.11**		
Proposed method	zebra crossings	**94.34**	90.09	**93.90**	**90.45**	
	bus stations	**99.59**	91.30	**100**		
	roadside parking spaces	77.44	80.95	78.14		

Note: The best results are in bold type.

To verify the effectiveness of the proposed method, ablation experiments were conducted on the road traffic multielement datasets. Such experiments compared the combination of k-means, mosaic data augmentation, and other attention mechanisms (such as the SE [49] attention mechanism, the convolutional block attention module (CBAM) [50] attention mechanism, and fusing the attention mechanisms into the same layer network structure as the ECA mechanism), by calculating their AP, precision, recall, and mAP values, as shown in Table 3.

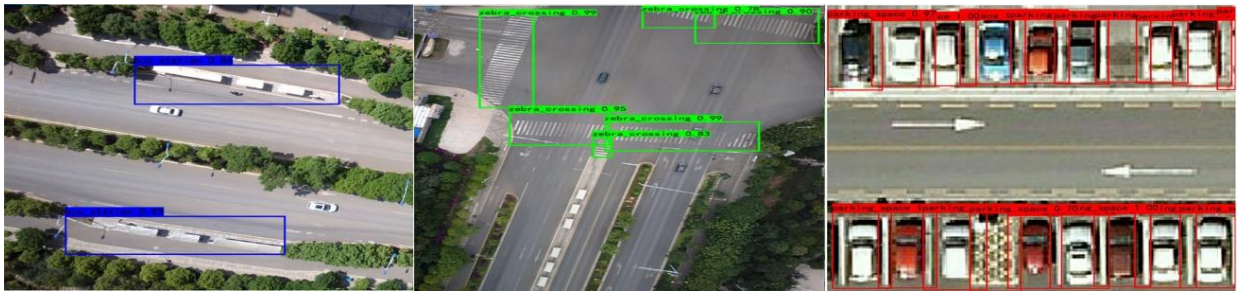
Table 3. Results of ablation tests.

Network Model	Transport Elements	AP	Precision	Recall	mAP	Rise Points
YOLOv4+k-means	zebra crossings	85.12	85.58	81.42	80.98	9.47
	bus stations	83.52	88.89	77.42		
	roadside parking spaces	74.30	77.94	75.62		
YOLOv4+mosaic	zebra crossings	88.31	87.95	87.17	82.44	8.01
	bus stations	84.93	**96.00**	77.42		
	roadside parking spaces	74.08	74.85	75.39		
YOLOv4+SE	zebra crossings	91.45	**91.43**	90.14	84.54	5.91
	bus stations	81.00	79.31	82.14		
	roadside parking spaces	**81.17**	**85.06**	**80.7**		
YOLOv4+CBAM	zebra crossings	92.07	89.63	90.95	86.92	3.53
	bus stations	92.01	92.00	88.46		
	roadside parking spaces	76.68	81.62	77.47		
Proposed method	zebra crossings	**94.34**	90.09	**93.90**	**90.45**	
	bus stations	**99.59**	91.30	**100**		
	roadside parking spaces	77.44	80.95	78.14		

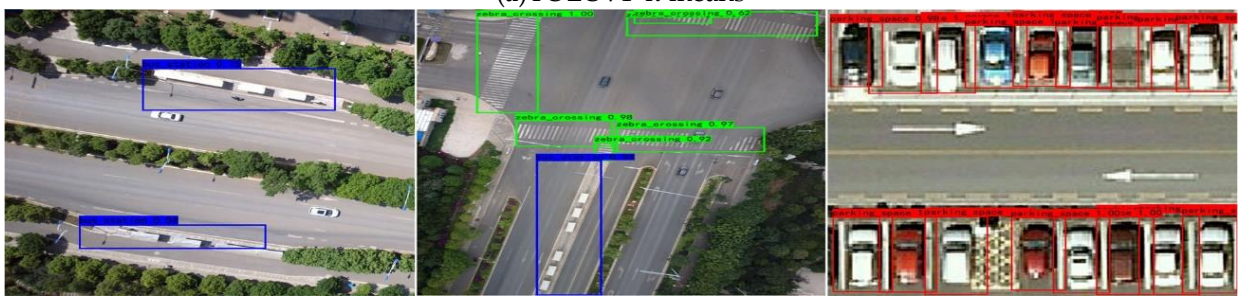
Note: The best results are in bold type.

To verify the practicality and effectiveness of the presented method in this paper, the UAV image map of small scenarios and the image map of large complex scenarios were selected for prediction experiments. The prediction results of the ablation experiment for small scenarios are shown in Figure 6. The prediction data are selected from the road traffic multielement datasets with several representative types of element scenarios, namely the single-element scenario, the multielement scenario, and the juxtaposed dense-element scenario. The single-element scenario contains only one type of traffic element, and the bus stations were selected as the detection object in the single-element scenario. The multielement scenario includes zebra crossings, bus stations, and roadside parking spaces. The juxtaposed dense-element scenario involves the detection and recognition of roadside parking spaces. According to the corresponding statistics, there are 2 bus stations in the single-element scenario; 4 zebra crossings, 3 roadside parking spaces, and 1 bus station in the multielement scenario; and 17 roadside parking spaces in the juxtaposed dense-element scenario. The predicted results of the ablation experiments in small scenarios are shown in Table 4.

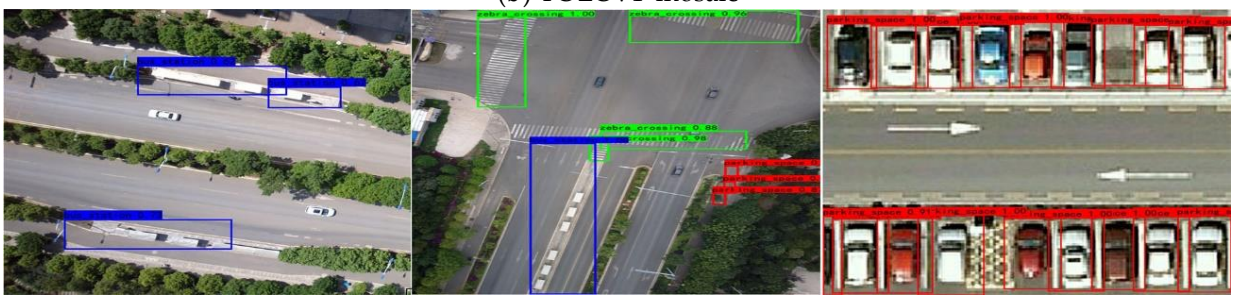
Combined with the predicted results in Figure 6 and Table 4, it is clear that the use of k-means clustering or mosaic data augmentation alone for the detection of multiple elements of road traffic suffers from leakage, proving that improving the algorithm from one side alone does not lead to a large improvement in the experimental results. Combined with the analysis of the ablation experiment detection results, the mosaic data augmentation method has the worst detection accuracy of only 74.08% for roadside parking spaces, followed by the k-means clustering method, as confirmed in the prediction results in Figure 6 and Table 4 which show missed detections in the prediction results. In terms of the overall prediction results, the detection results improve with the addition of the attention mechanism, and the proposed method has the highest number of optimal detections at four. In particular, the detection of zebra crossings and roadside parking spaces reaches 98.25% and 99.88% for the detection of multiple elements and dense side-by-side scenarios, respectively. Although the detection of bus stations in complex scenarios with the addition of the SE attention mechanism achieves the best detection, the detection accuracy of the proposed method reaches 98%, which is only 2% different from the detection method with the addition of the SE attention mechanism.



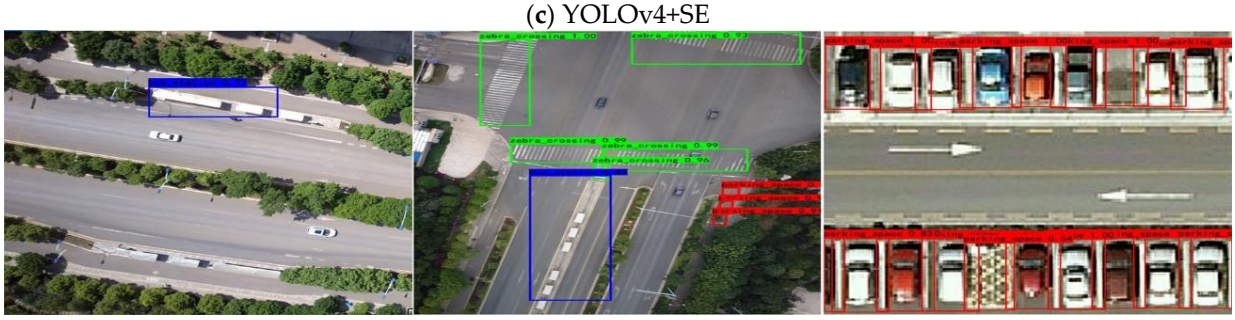
(**a**)YOLOv4+k-means



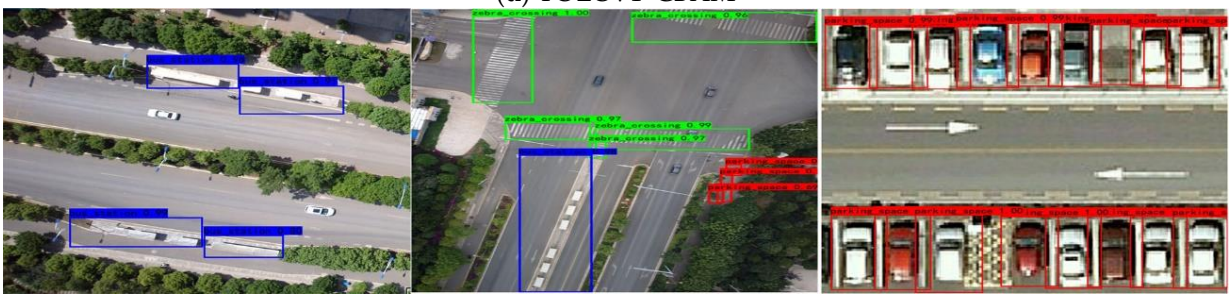
(**b**) YOLOv4+mosaic



(**c**) YOLOv4+SE



(**d**) YOLOv4+CBAM



(**e**) Proposed method

Figure 6. Plot of predicted results of ablation experiments for small scenarios, from left to right, single-element scenario, multielement scenario, and juxtaposed dense-element scenario. (**a**) YOLOv4 with k-means; (**b**) YOLOv4 with mosaic; (**c**) YOLOv4 with SE attention mechanism; (**d**) YOLOv4 with CBAM attention mechanism; (**e**) proposed method.

Table 4. Predicted results of ablation experiments in small scenarios.

Network Model	Transport Elements	Single-Element		Multielement		Juxtaposed Dense-Element	
		Number	AP	Number	AP	Number	AP
YOLOv4+k-means	zebra crossings	-	-	4	86.50	-	-
	bus stations	2	89.50	0	Leakage	-	-
	roadside parking spaces	-	-	0	Leakage	17	95.71
YOLOv4+mosaic	zebra crossings	-	-	4	95.50	-	-
	bus stations	2	64.50	1	94.00	-	-
	roadside parking spaces	-	-	0	Leakage	17	**99.88**
YOLOv4+SE	zebra crossings	-	-	4	95.50	-	-
	bus stations	2	67.50	1	**100**	-	-
	roadside parking spaces	-	-	3	87.33	17	99.47
YOLOv4+CBAM	zebra crossings	-	-	4	96.75	-	-
	bus stations	1	77.00	1	82.00	-	-
	roadside parking spaces	-	-	3	87.68	17	96.29
Proposed method	zebra crossings	-	-	4	**98.25**	-	-
	bus stations	2	**98.50**	1	98.00	-	-
	roadside parking spaces	-	-	3	**88.67**	17	**99.88**

Note: "-" in the table indicates that the images measured do not contain this category, and the bolded font is the best result for each. Number is the number of road traffic elements that have been correctly detected.

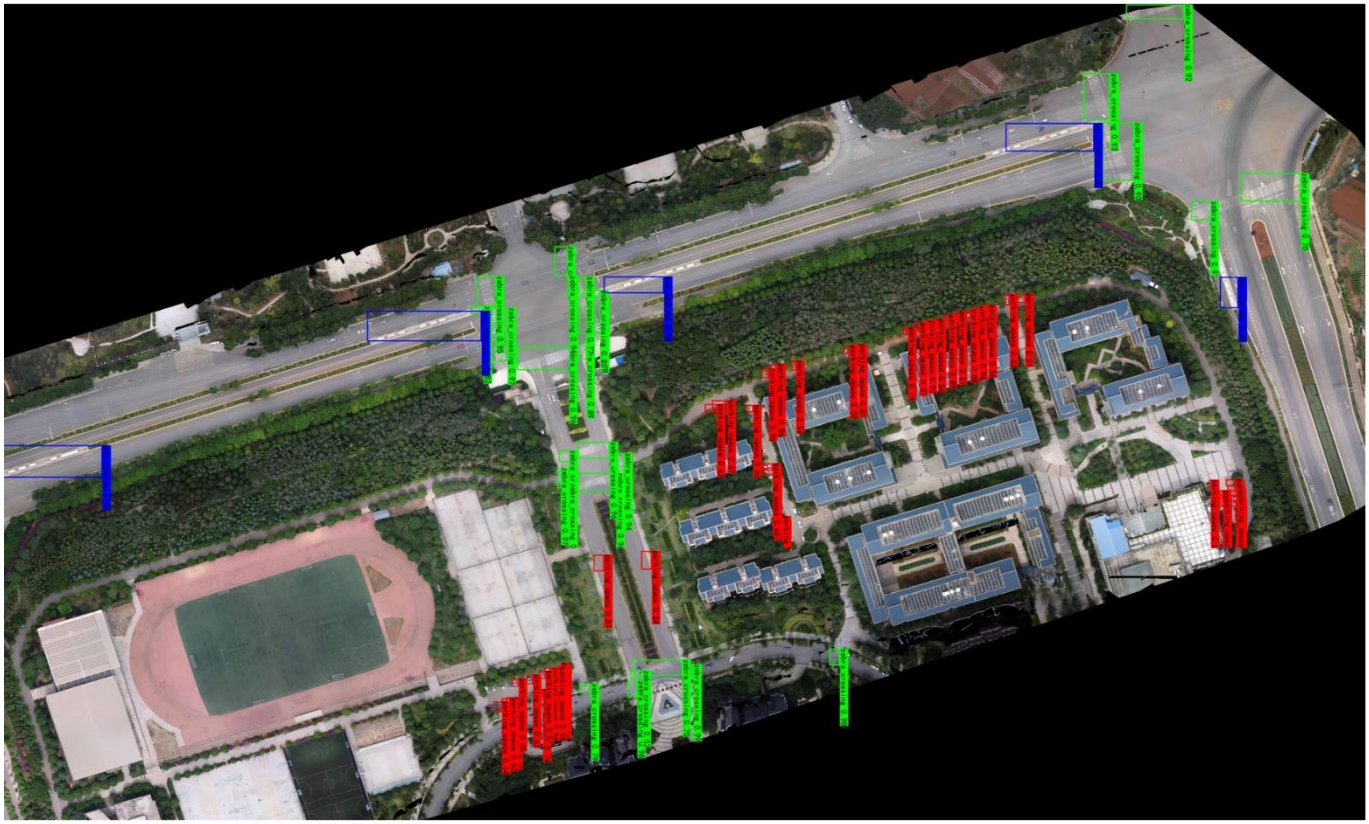
The predicted results of the ablation experiment in a large complex scene are shown in Figure 7. The large scene map is an ortho mosaic image generated from the remote sensing image captured by the UAV and processed by Pix4D software, which covers a total area of 302,813 m^2. Upon counting, it is determined that the large complex scene includes 18 zebra crossings, 5 bus stations, and 58 roadside parking spaces. The specific prediction results are shown in Table 5.

Table 5. Predicted results of ablation experiments in large complex scenarios.

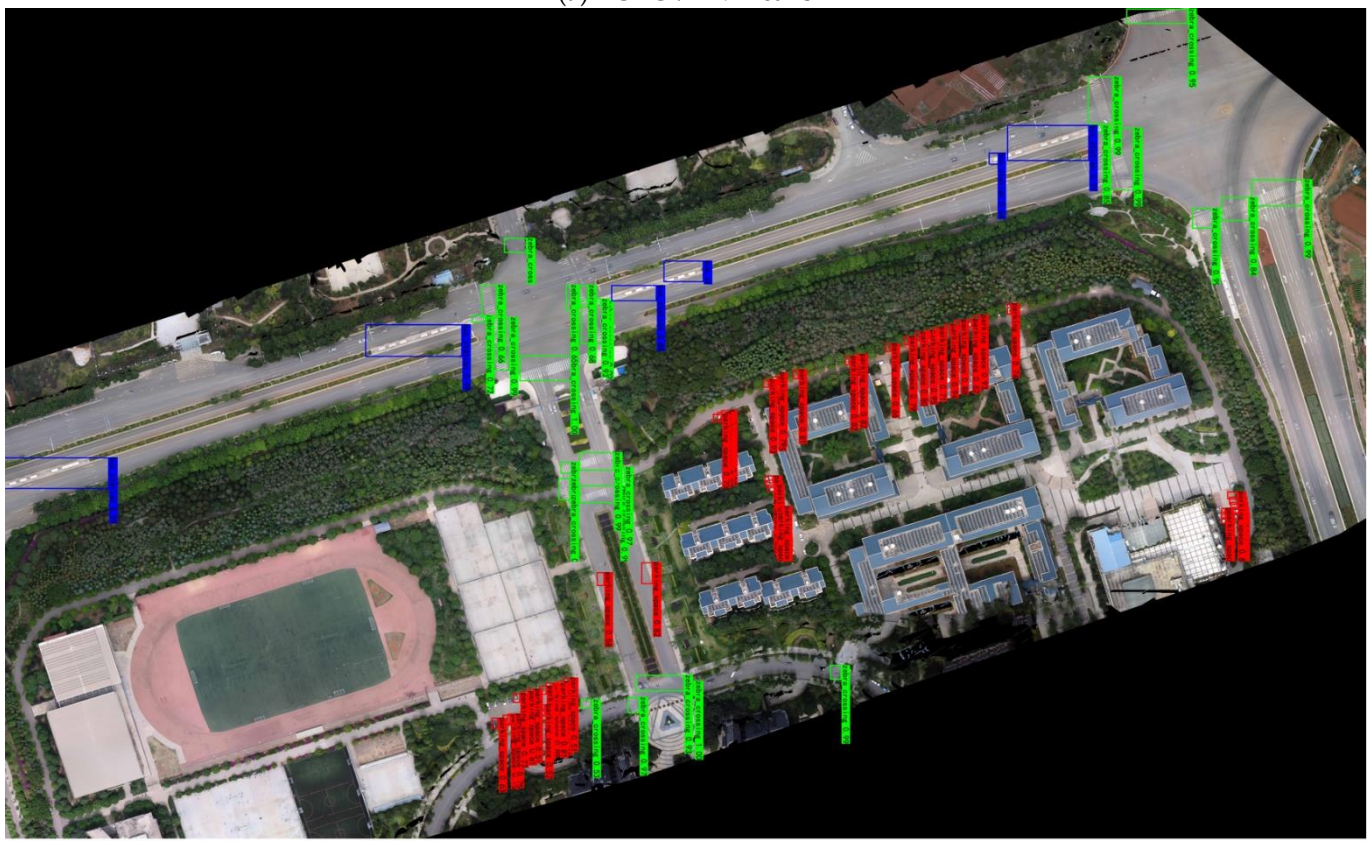
Network Model	Transport Elements	Number	AP
YOLOv4+k-means	zebra crossings	17	89.26
	bus stations	4	85.75
	roadside parking spaces	42	72.49
YOLOv4+mosaic	zebra crossings	18	89.39
	bus stations	4	86.00
	roadside parking spaces	54	77.02
YOLOv4+SE	zebra crossings	18	89.48
	bus stations	5	85.11
	roadside parking spaces	56	**78.23**
YOLOv4+CBAM	zebra crossings	17	92.14
	bus stations	5	89.33
	roadside parking spaces	43	73.03
Proposed method	zebra crossings	18	**92.17**
	bus stations	5	**93.40**
	roadside parking spaces	48	76.48

Note: The best results are in bold type. Number is the number of road traffic elements that have been correctly detected.

Combined with the prediction results in Figure 7 and Table 5, it is clear that several of the above algorithms miss detections in large complex scenarios, especially when detecting roadside parking spaces. The reason for this is that for large images, roadside parking comprises a small target detection, and most roadside parking spaces are covered by greenery; thus, the feature information is not obvious, resulting in missed detection. For the detection of other objects, the algorithm in this paper can still exhibit good results. The average detection accuracy of zebra crossings and bus stations in large complex scenarios can reach 92.17% and 93.40%, respectively, corresponding to the best result in the ablation experiment.

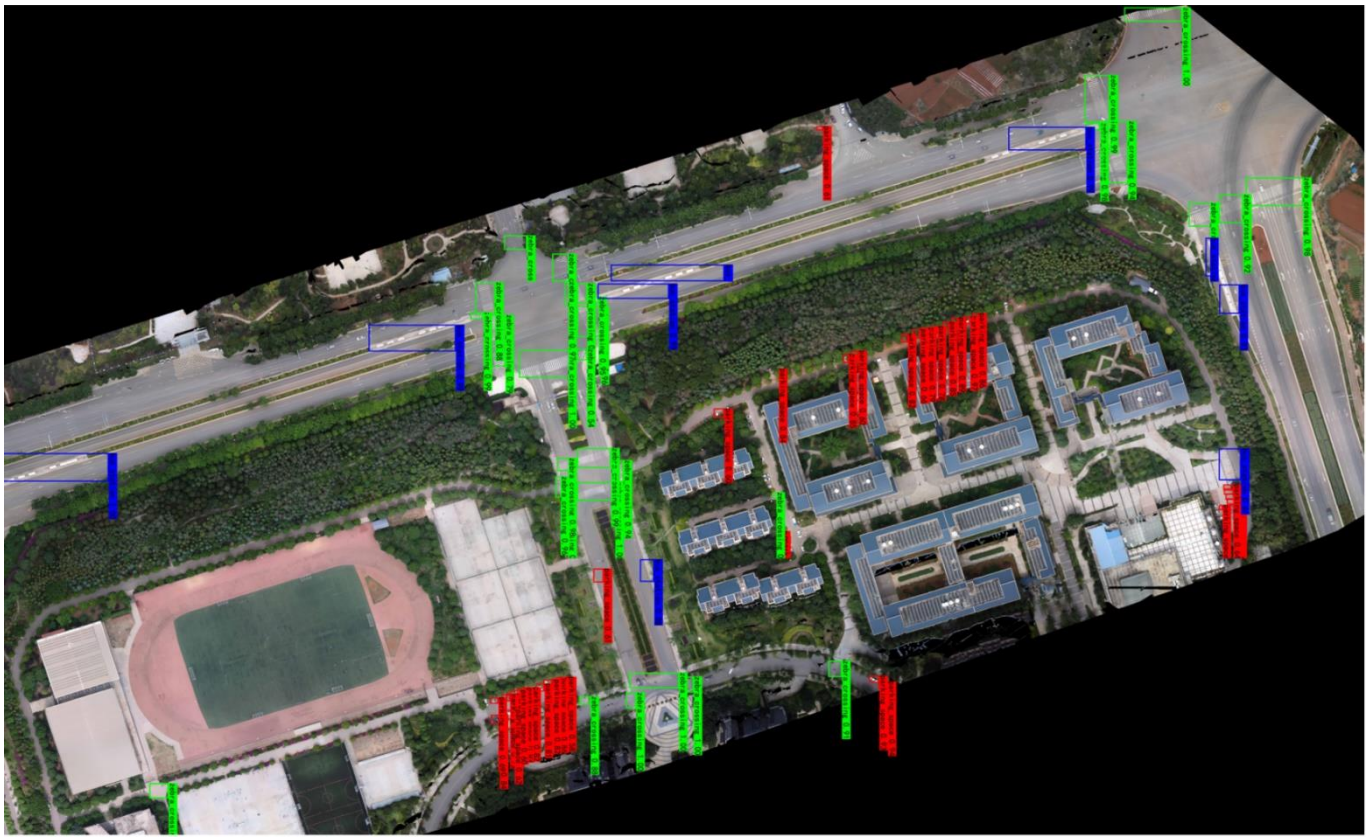


(**a**) YOLOv4+k-means

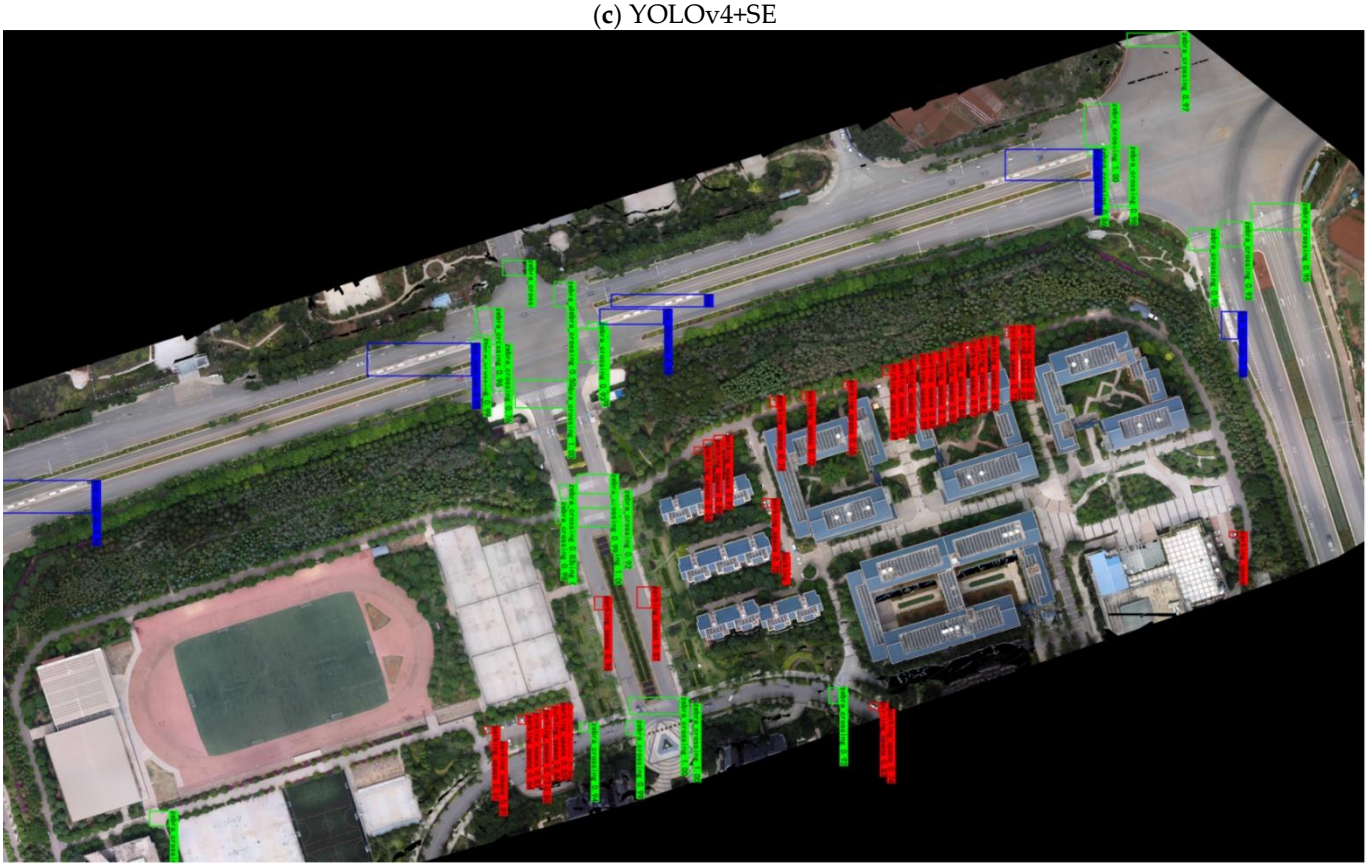


(**b**) YOLOv4+mosaic

Figure 7. *Cont.*

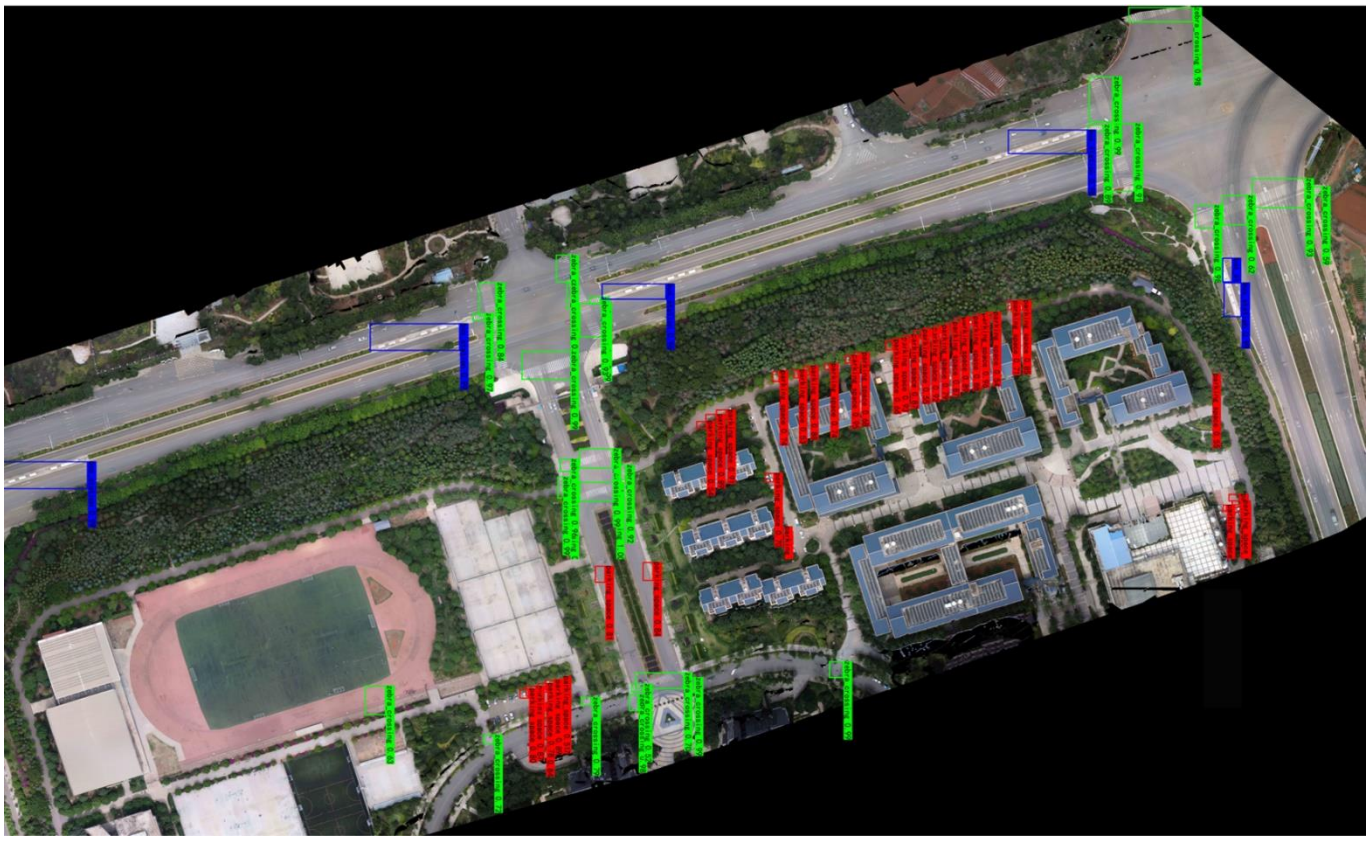


(**c**) YOLOv4+SE



(**d**) YOLOv4+CBAM

Figure 7. *Cont*.



(**e**) Proposed method

Figure 7. Predicted results of the ablation experiment in a large and complex scenario. (**a**) YOLOv4 with k-means; (**b**) YOLOv4 with mosaic; (**c**) YOLOv4 with SE attention mechanism; (**d**) YOLOv4 with CBAM attention mechanism; (**e**) proposed method.

5. Discussion

From the above experimental results and analysis, we find that the present method exhibits a large improvement in the mAP compared to several other methods, with the increase points ranging from 3.53% to 36.51%, verifying that the YOLOv4 model incorporating the ECA mechanism presented in this paper can effectively improve the road traffic multielement detection accuracy. Consistent with the results of previous studies, the experimental results of combining other dominant attention mechanism modules in the same network location are improved compared to the original YOLOv4 network; however, the improvement is not as good as the present method, indicating that the fused attention mechanism has a positive effect on the network training model. The proposed YOLOv4 algorithm with the fused ECA mechanism is the best. This demonstrates the practicality and superiority of the proposed method, which can be directly applied to image maps in large scenarios and provides a more intelligent and convenient method for updating the basic traffic geographic information database. Moreover, the proposed method still achieves better results than several other methods in complex large scenes, which proves its practicality and superiority. The proposed method can be directly applied to image maps in large scenarios, thus providing a more intelligent and convenient method for updating geographic information database. However, in large complex scenarios, there is still a missing detection phenomenon for roadside parking spaces. This is because in this scenario, roadside parking spaces are easily obscured, and they are small targets for detection. This problem requires subsequent research on how to improve the detection of small targets in large complex scenarios and better extract feature information from small targets.

6. Conclusions

To address the problems of low data extraction, poor automation, and high demand for traffic element information, an automatic recognition and detection method based on YOLOv4 multiple road traffic elements combined with an attention mechanism based on UAV remote sensing images is proposed in this paper. The method achieves 90.45% mAP in the detection of multiple road traffic elements, which is 18.80% better than the original YOLOv4 network. The experimental results verify that the method in this paper provides a new idea for updating and improving the basic traffic geographic information database.

However, the method in this paper also has shortcomings. The experiment focuses only on zebra crossings, bus stations, and roadside parking spaces, and the subsequent work will expand the datasets to complete automatic identification and detection of more elements.

Author Contributions: L.H.: funding acquisition, directing, project administration, and manuscript review. M.Q.: algorithm proposed and testing, data processing, and manuscript writing. A.X.: research conceptualization. Y.S. and J.Z.: dataset production. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, "Study on the slope gradient effect of land distribution and change of urban construction in southern mountainous areas" (No. 41961039), and the Applied Basic Research Program of Yunnan Province, "Research on Remote Sensing Image Matching Based on Image Transformation and Deep Feature Learning" (No. 202101AT070102).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data have not been made publicly available, but it can be used for scientific research. Other researchers can send emails to the first author if needed.

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

References

1. Xiao, J. Design of urban basic traffic management system based on geographic information system. In Proceedings of the 29th Chinese Control Conference, Beijing, China, 29–31 July 2010.
2. Wang, F.; Wang, J.; Li, B.; Wang, B. Deep attribute learning based traffic sign detection. *Jilin Daxue Xuebao (Gongxueban)* **2018**, *48*, 319–329.
3. Li, H.J.; Sun, F.M.; Liu, L.J.; Wang, L. A novel traffic sign detection method via color segmentation and robust shape matching. *Neurocomputing* **2015**, *169*, 77–88. [CrossRef]
4. Zhao, Y.W. Image processing based road traffic sign detection and recognition method. *Transp. World* **2018**, *2018*, 42–43.
5. Zhang, S.F.; Zhu, X.Y.; Lei, Z.; Shi, H.L.; Wang, X.B.; Li, S.Z. Faceboxes: A cpu real-time face detector with high accuracy. In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017.
6. Bagschik, G.; Menzel, T.; Maurer, M. Ontology based scene creation for the development of automated vehicles. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium, Suzhou, China, 26–30 September 2018.
7. Liao, M.H.; Shi, B.G.; Bai, X.; Wang, X.G.; Liu, W.Y. TextBoxes: A fast text detector with a single deep neural network. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–10 February 2017.
8. Meng, Y.B.; Shi, D.W.; Liu, G.H.; Xu, S.J.; Jin, D. Dense irregular text detection based on multi-dimensional convolution fusion. *Guangxue Jingmi Gongcheng* **2021**, *29*, 2210–2221. [CrossRef]
9. Berkaya, S.K.; Gunduz, H.; Ozsen, O.; Akinlar, C.; Cunal, S. On circular traffic sign detection and recognition. *Expert Syst. Appl.* **2016**, *48*, 67–75. [CrossRef]
10. Shi, X.P.; He, W.; Han, L.Q. A road edge detection algorithm based on the hough transform. *Trans. Intell. Syst.* **2012**, *7*, 81–85.
11. He, J.P.; Ma, Y. Triangle Traffic sign detection approach based on shape information. *Comput. Eng.* **2010**, *36*, 198–199+202.
12. Creusen, I.M.; Wijnhoven, R.; Herbschleb, E. Color exploitation in hog-based traffic sign detection. In Proceedings of the IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010.
13. Hao, X.; Zhang, G.; Ma, S. Deep learning. *Int. J. Semant. Comput.* **2016**, *10*, 417–439. [CrossRef]
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y. SSD: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.

16. 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.
17. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. Available online: https://arxiv.org/abs/1804.02767 (accessed on 8 April 2018).
18. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal Speed and Accuracy of Object Detection. Available online: https://arxiv.org/abs/2004.10934 (accessed on 23 April 2020).
19. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding Yolo Series in 2021. Available online: https://arxiv.org/abs/2107.08430 (accessed on 6 August 2021).
20. Girshick, R.; Donahue, J.; Darrell, T.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
21. Girshick, R. Fast r-cnn. In Proceedings of the 15th IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015.
22. Ren, S.Q.; He, K.M.; 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]
23. Shan, H.; Zhu, W. A small traffic sign detection algorithm based on modified ssd. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *646*, 012006. [CrossRef]
24. Chen, P.D.; Huang, L.; Xia, Y.; Yu, X.N.; Gao, X.X. Detection and recognition of road traffic signs in uav images based on mask r-cnn. *Remote Sens. Land Resour.* **2020**, *32*, 61–67.
25. Lodhi, A.; Singhal, S.; Massoudi, M. Car traffic sign recognizer using convolutional neural network cnn. In Proceedings of the 6th International Conference on Inventive Computation Technologies, Coimbatore, India, 20–22 January 2021.
26. Guo, J.K.; Lu, J.Y.; Qu, Y.Y.; Li, C.H. Traffic-sign spotting in the wild via deep features. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium, Suzhou, China, 26–30 September 2018.
27. Jin, Z.Z.; Zheng, Y.F. Research on application of improved yolov3 algorithm in road target detection. *J. Phys. Conf. Ser.* **2020**, *1654*, 012060.
28. Wang, Q.L.; Wu, B.G.; Zhu, P.F.; Li, P.H.; Zuo, W.M.; Hu, Q.H. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, Online, 14–19 June 2020.
29. Cgvict. RoLabelImg. Available online: https://github.com/cgvict/roLabelImg (accessed on 23 June 2020).
30. Zheng, Z.H.; Wang, P.; Liu, W.; Li, J.Z.; Ye, R.G.; Ren, D.W. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
31. Liu, X.F.; Guan, Z.W.; Song, Y.Q.; Chen, D.S. An optimization model of UAV route planning for road segment surveillance. *J. Cent. South Univ.* **2014**, *21*, 2501–2510. [CrossRef]
32. Cheng, L.H.; Zhong, L.; Tian, S.S.; Xing, J.X. Task assignment algorithm for road patrol by multiple uavs with multiple bases and rechargeable endurance. *IEEE Access* **2019**, *7*, 144381–144397. [CrossRef]
33. Elloumi, M.; Dhaou, R.; Escrig, B.; Idoudi, H.; Saidane, H. Monitoring road traffic with a uav-based system. In Proceedings of the 2018 IEEE Wireless Communications and Networking Conference, Barcelona, Spain, 15–18 April 2018.
34. Yang, J.; Zhang, J.L.; Ye, F.; Cheng, X.H. A uav based multi-object detection scheme to enhance road condition monitoring and control for future smart transportation. In Proceedings of the 1st EAI International Conference on Artificial Intelligence for Communications and Networks, Harbin, China, 25–26 May 2019.
35. Wang, X.; Ouyang, C.T.; Shao, X.L.; Xu, H. A method for uav monitoring road conditions in dangerous environment. *J. Phys. Conf. Ser.* **2021**, *1792*, 012050. [CrossRef]
36. Huang, H.L.; Savkin, A.V.; Huang, C. Decentralised autonomous navigation of a uav network for road traffic monitoring. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 2558–2564. [CrossRef]
37. Liu, X.F.; Peng, Z.R.; Zhang, L.Y. Real-time uav rerouting for traffic monitoring with decomposition based multi-objective optimization. *J. Intell. Robot. Syst. Theor. Appl.* **2019**, *94*, 491–501. [CrossRef]
38. Pan, Y.F.; Zhang, X.F.; Cervone, G.; Yang, L.P. Detection of asphalt pavement potholes and cracks based on the unmanned aerial vehicle multispectral imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3701–3712. [CrossRef]
39. Saad, A.M.; Tahar, K.N. Identification of rut and pothole by using multirotor unmanned aerial vehicle (UAV). *Measurement* **2019**, *137*, 647–654. [CrossRef]
40. Roberts, R.; Inzerillo, L.; Mino, G.D. Using uav based 3d modelling to provide smart monitoring of road pavement conditions. *Information* **2020**, *11*, 568. [CrossRef]
41. Wang, S.J.; Jiang, F.; Zhang, B.; Ma, R.; Hao, Q. Development of uav-based target tracking and recognition systems. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 3409–3422. [CrossRef]
42. Liu, H.; Mu, C.P.; Yang, R.X.; He, Y.; Wu, N. Research on object detection algorithm based on uva aerial image. In Proceedings of the 7th IEEE International Conference on Network Intelligence and Digital Content, Beijing, China, 17–19 November 2021.
43. Pelleg, D.; Moore, A. X-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, USA, 29 June–2 July 2000.

44. He, K.M.; Zhang, X.Y.; Ren, X.Q.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
45. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
46. Liu, S.; Qi, L.; Qin, H.F.; Shi, J.P.; Jia, J.Y. Path aggregation network for instance segmentatio. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
47. Wang, C.Y.; Yuan, H.; Wu, Y.H.; Chen, P.Y. CSPNet: A new backbone that can enhance learning capability of cnn. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, Online, 14–19 June 2020.
48. Everingham, M.; Eslami, S.; Gool, L.V.; Williams, C.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
49. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E.H. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
50. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.