future internet

MDPI

Article

SMYOLO: Lightweight Pedestrian Target Detection Algorithm in Low-Altitude Scenarios

Weiwei Zhang [1,2,*], Xin Ma [1,2], Yuzhao Zhang [1], Ming Ji [1,2] and Chenghui Zhen [1,2]

[1] College of Engineering, Huaqiao University, Quanzhou 362021, China; xinma@stu.hqu.edu.cn (X.M.); ZYZ@hqu.edu.cn(Y.Z.); jiming@stu.hqu.edu.cn (M.J.); chzhen@stu.hqu.edu.cn (C.Z.)

[2] Fujian Provincial Academic Engineering Research Centre in Industrial Intellectual Techniques and Systems, Quanzhou 362021, China

* Correspondence: weiweizh@hqu.edu.cn

Abstract: Due to the arbitrariness of the drone's shooting angle of view and camera movement and the limited computing power of the drone platform, pedestrian detection in the drone scene poses a greater challenge. This paper proposes a new convolutional neural network structure, SMYOLO, which achieves the balance of accuracy and speed from three aspects: (1) By combining deep separable convolution and point convolution and replacing the activation function, the calculation amount and parameters of the original network are reduced; (2) by adding a batch normalization (BN) layer, SMYOLO accelerates the convergence and improves the generalization ability; and (3) through scale matching, reduces the feature loss of the original network. Compared with the original network model, SMYOLO reduces the accuracy of the model by only 4.36%, the model size is reduced by 76.90%, the inference speed is increased by 43.29%, and the detection target is accelerated by 33.33%, achieving minimization of the network model volume while ensuring the detection accuracy of the model.

Keywords: model compression; pedestrian detection; deep learning; drone scene

Citation: Zhang, W.; Ma, X.; Zhang, Y.; Ji, M.; Zhen, C. SMYOLO: Lightweight Pedestrian Target Detection Algorithm in Low-Altitude Scenarios. *Future Internet* **2022**, *14*, 21. https://doi.org/10.3390/fi14010021

Academic Editor: Eirini Eleni Tsiropoulou

Received: 2 December 2021
Accepted: 29 December 2021
Published: 4 January 2022

1. Introduction

Unmanned aerial vehicles (UAV) have played an important role in traffic monitoring, route inspection, military, forestry [1–4], and other fields in recent years. Pedestrian detection is becoming more and more important in applications such as intelligent monitoring, re-identification of people, and autonomous driving. Therefore, drones to detect pedestrians have become an emerging technology. Unlike ordinary target detection, pedestrian detection based on low-altitude drones can provide more comprehensive information and lay the foundation for an in-depth understanding and analysis of pedestrians. However, due to the special viewing angle of low-altitude drones and the low computing power of embedded devices, quickly and accurately detecting pedestrians is still a challenging problem.

Traditional target detection algorithms lack effective image feature representation methods. The model's generalization ability is insufficient, and it is challenging to complete target detection through general abstract features [5]. The target detection algorithm based on deep learning uses a convolutional neural network (CNN) to extract the target feature richly to complete target detection. Although the current target detection methods based on deep learning have achieved specific achievements, there are still significant challenges in pedestrian detection in unique low-altitude drone scenes. These challenges are mainly: (i) Due to the instability and randomness of the drone shooting camera, it will cause camera shake and aspect ratio changes, which will complicate the background and increase the difficulty of detection to a certain extent; (ii) for pedestrian detection, the drone will distort pedestrians and easily cause false detection when performing overhead shooting; (iii) the scale mismatch between the dataset in the pre-training weight used in the training process

and the small target dataset will cause the loss of object feature representation, which will reduce the performance of the detector; (iv) the limited computing power of the drone's airborne platform poses a more significant challenge to the complexity of the detection algorithm; and (v) the latency of high-precision target models is high, affecting how to balance detection accuracy and real-time performance.

In response to the above problems, this article proposes a new convolutional neural network structure SMYOLO.

- Aiming at the problem of limited computing power on the drone's airborne platform and high latency of high-precision target models, the SMYOLO target detection model is proposed. The conventional convolution is decomposed and replaced with deep separable convolution and dot convolution, and the activation function is replaced, so as to compress the YOLOv4 model and realize the deployment on the embedded platform.
- Because of the reduced model accuracy after the compressed model and the problem of false pedestrian detection in pedestrian detection, the pre-training dataset and the training dataset are scaled to match, which will improve the feature representation ability of the detector during the training process. SMYOLO has added the BN layer to improve the generalization ability of the detector, thereby improving the accuracy of SMYOLO.

This algorithm reduces the model's size while ensuring the computing power of the drone's airborne platform, which is very important for the real-time detection of pedestrian scenes. The experimental results and analysis show that compared with the current mainstream methods, the method in this paper increases the detection speed while maintaining the detection accuracy and ensures the detection accuracy of the model while minimizing the volume of the network model as much as possible.

2. Related Work

2.1. Research on High-Precision Target Detection Algorithms

With the development of neural networks, target detection algorithms based on deep learning have made certain achievements. Representative algorithms are: RCNN(Region-based CNN) series (RCNN [6], Fast RCNN [7] and Faster RCNN [8]), SSD(Single Shot MultiBox Detector) [9] and YOLO(You Only Look Once) series (YOLO [10], YOLOV2 [11], YOLOV3 [12], YOLOV4 [13]). This is more research on pedestrian detection algorithms in ordinary scenes, but there are fewer pedestrian detection algorithms in drone scenes. Although pedestrians are detected in ordinary scenes and drone scenes, there are certain differences between pedestrians in the two scenes. For example, the pedestrian target in the drone scene is relatively small, which makes detection difficult. Liu [14] improved the YOLOv3 network structure. By increasing the operation of the convolutional layer in the early stage to enrich the spatial information, the performance of small target detection was significantly improved. Yu [15] proposed a scale matching method to keep the feature distribution of the network pre-training dataset and the detector dataset consistent and improve the quality of the small target detection model by improving the similarity of the data distribution. Shao [16] used the k-means clustering algorithm to extract pedestrian areas, used RPN to generate recommended candidate areas quickly, and expanded the structure of Faster-RCNN through the method of high- and low-level feature fusion to improve the detection ability of small pedestrians.

2.2. Research on Lightweight Target Detection Algorithms

Embedded devices have relatively small computing capabilities. The deployment of a complete target detection model on embedded devices will lead to problems such as high energy consumption, high computational load, and high latency, making it hard to meet real-time performance. In this case, the demand for lightweight networks has grown. Wu [17] pruned the trained model through the channel pruning algorithm of YOLOv4, simplifying the structure and parameters of the model under the premise of ensuring

accuracy and realizing real-time detection. Yu [18] used the focal loss to optimize the loss function to improve the accuracy and used a pruning algorithm to simplify the network to speed up the detectionẎan [19] realized the deployment of the model on the embedded platform by replacing the activation function of YOLOv4 with the ELU activation function and adding the SE attention module to the backbone. Zhao [20] proposed that based on YOLO-LITE as the backbone network, mixed YOLOv3-LITE supplements residual block (ResBlocks) and parallel high-to-low resolution subnetworks. ALFASLY [21] proposed to adaptively zoomin the input frame by splitting it into nonoverlapped tiles and paying attention only to the important tiles. It has obtained a fully convolutional pedestrian detection model that can be run on low computational resources. Ke [22] processed with a nonoverlapped image blocking data augmentation method, and then input them into the YOLOv3 detector to obtain the object position information. An LCNN-based pedestrian re-identification model is used to extract the features of the object.

2.3. High-Precision Target Detection Algorithm: YOLOv4

YOLOv4 is a first-level object detector, an improved target detection algorithm of YOLOv3. The architecture of YOLOv4 is CSPDarkNet53+SPP+PANET+YOLOv3, and the main structure is shown in Figure 1. The head is the same as YOLOv3 and compared with YOLOv3, YOLOv4 introduces CSPNet, which changes its backbone structure to CSPDarkNET53, containing 29 convolutions. The receptive field is 725 × 725, and the parameter is 27.6 million. Compared with Darknet-53, this backbone structure uses the CSP module to first divide the feature map of the base layer into two parts and then merge through the cross-stage hierarchical structure to enhance the learning ability of the CNN and ensure accuracy while reducing the amount of calculation. In order to increase the receptive field, YOLOv4 uses PANet [23] instead of FPN [24] for the parameter aggregation to be suitable for target detection of different scales. It uses spatial pyramid pooling (SPP) [25] as an additional module of Neck to perform multi-scale fusion and improve feature extraction capabilities. Finally, to train the network more efficiently on a single GPU, YOLOv4 uses data augmentation methods to mosaic and spontaneous training in the network and uses genetic algorithms to select the best hyperparameters. Through an analysis, the speed of the YOLOv4 algorithm can significantly meet the accuracy requirements of the pedestrian detection task, but due to the poor computing power of the embedded device, it cannot meet the real-time requirements. Therefore, this paper improves the YOLOv4 network to enable real-time pedestrian detection in the low-altitude UAV scene.

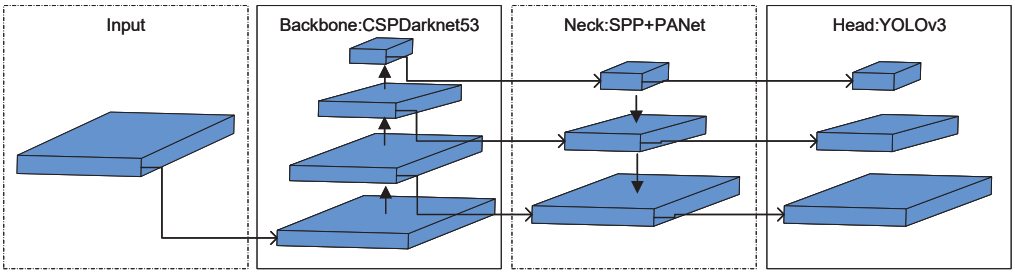


Figure 1. YOLOv4 structure diagram.

3. Proposed Method

The flow of the low-altitude pedestrian detection algorithm proposed in this paper is shown in Figure 2. First, by improving the YOLOv4 model, a new target detection model is proposed: SMYOLO. SMYOLO uses deep separable convolution to extract features and changes the activation function to a lightweight activation function, which will improve the computational efficiency of the convolutional network and reduce the number of parameters, thereby improving the detection speed of the convolutional network model, so that it can perform real-time detection on airborne equipment with less computing power. Secondly, SMYOLO extracts the pedestrian data in the VisDrone dataset to generate a

pedestrian dataset, and use the method of scale matching to increase the feature representation ability of the detector to improve the detection accuracy of the SMYOLO convolutional network model to achieve the balance between SMYOLO's accuracy and speed.

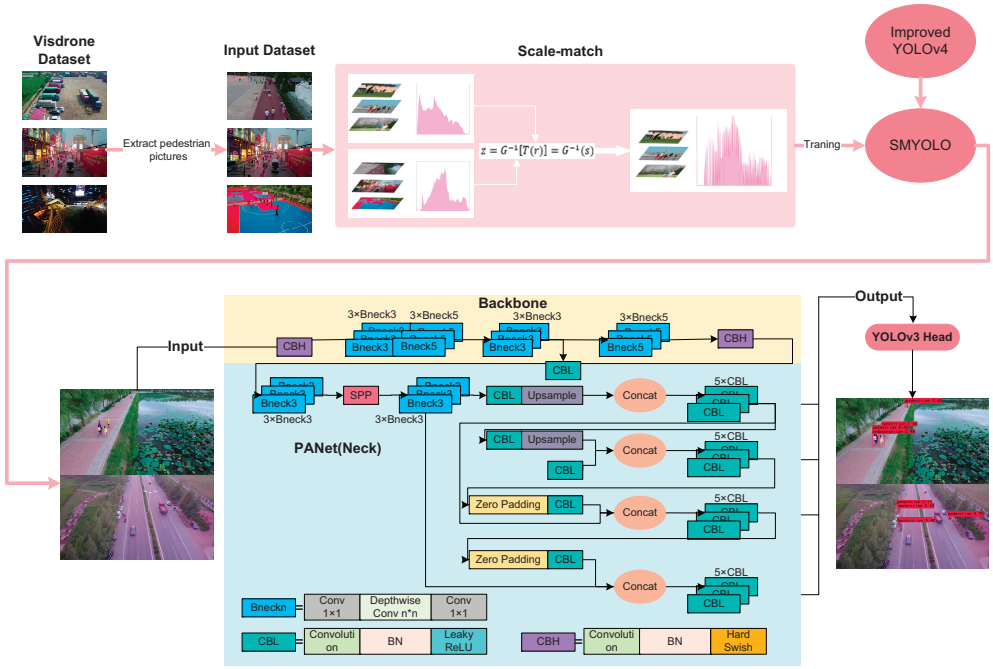


Figure 2. Low-altitude pedestrian detection algorithm flow chart.

3.1. Low-Altitude Pedestrian Detection Network Architecture: SMYOLO

The ideal detector must achieve high accuracy in positioning and recognition and high efficiency in terms of time. In recent years, many practical target detection algorithms have been proposed, such as SSD and Fast R-CNN. This article selects the recently proposed CNN architecture YOLOv4. Compared with other networks, this network achieves an optimal balance of speed and accuracy. However, if it is directly deployed on embedded devices, it will lead to the overload of airborne drone equipment. Therefore, the SMYOLO network architecture is proposed. The proposed SMYOLO network follows the YOLO target detection idea, using the characteristics of the picture to predict each frame, that is, predicting all the frames of all categories at one time so that the end-to-end prediction can guarantee very high detection speed.

3.1.1. Depth Separable Convolution and Point Convolution

SMYOLO is formed by stacking with depth separable convolution while using point convolution to increase the depth of the network structure, making the SMYOLO network "narrow and deep". In conventional convolution, assuming that the input feature dimension is (H_W, H_W, M), the convolution kernel is (N, H_X, H_X, M), and the output feature map is (H_K, H_K, M). The calculation process is shown in the Figure 3.

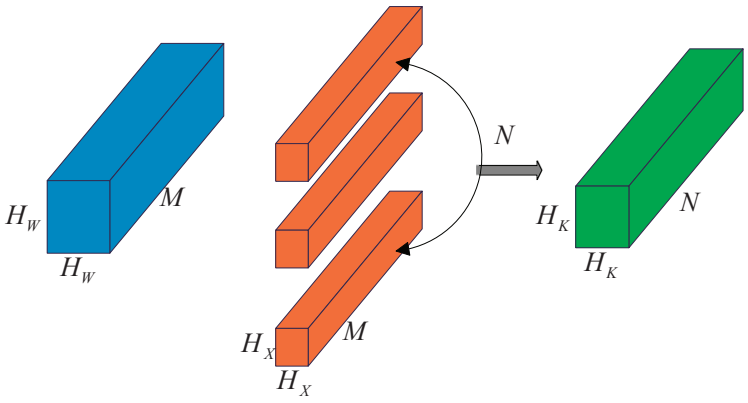


Figure 3. Conventional convolution calculation process.

The amount of calculation required for conventional convolution to complete this process is shown in Equation (1).

$$Calculation_{conv} = H_K \times H_K \times H_X \times H_X \times N \times M \tag{1}$$

In-depth separable convolution is composed of depth convolution (Depthwise) and point convolution (Pointwise). Deep convolution works on each convolution channel, and point convolution integrates the convolution output of each channel. The calculation process is shown in Figure 4

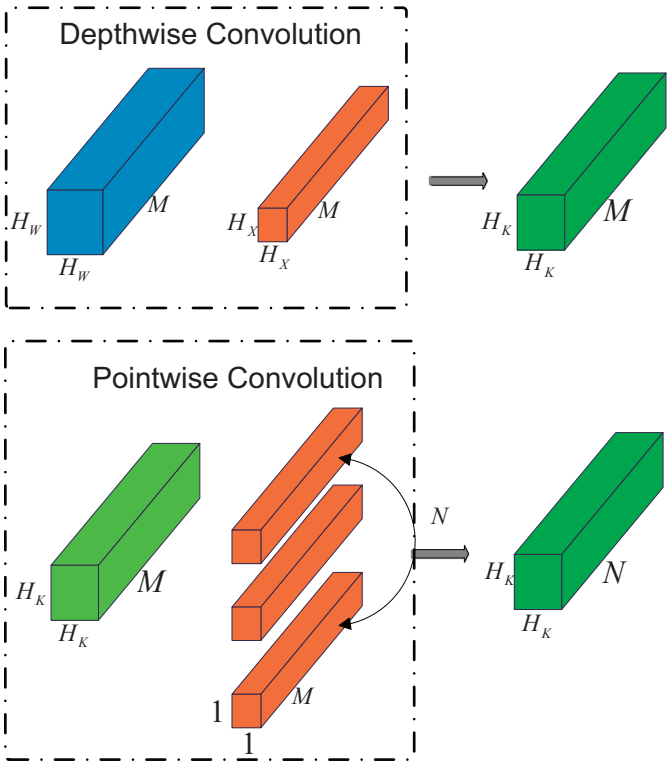


Figure 4. Depth separable convolution calculation process.

SMYOLO's depth separable convolution requires the amount of calculation to complete this process is shown in Equation (2).

$$Calculation_{dw_conv} = H_K \times H_K \times H_X \times H_X \times M + H_K \times H_K \times N \times M \tag{2}$$

The SMYOLO compression ratio is the ratio of the parameter amount of the conventional convolution to the parameter amount of the depth separable convolution, namely:

$$Ratio = \frac{H_K \times H_K \times H_X \times H_X \times M + H_K \times H_K \times N \times M}{H_K \times H_K \times H_X \times H_X \times N \times M} = \frac{1}{N} + \frac{1}{{H_X}^2} \quad (3)$$

It can be seen from the above formula that the parameter amount of the depth separable convolution is less than that of the conventional convolution, so replacing the conventional convolution with the depth separable convolution can reduce the amount of calculation to a certain extent. The basic structure of the deep separable convolutional network is shown in Figure 5. It will use deep convolution (Depthwise) and point convolution (Pointwise) instead of conventional convolution. In the SMYOLO target detection model, we not only replaced the conventional convolution in the head module of YOLOv4 with a depth separable convolution, but also modified the 3×3 conventional convolution in the PANet module in YOLOv4. We replaced it with depth separable convolution, which will reduce the number of parameters to a certain extent and thus reduce the amount of calculation deployed on drones' onboard equipment.

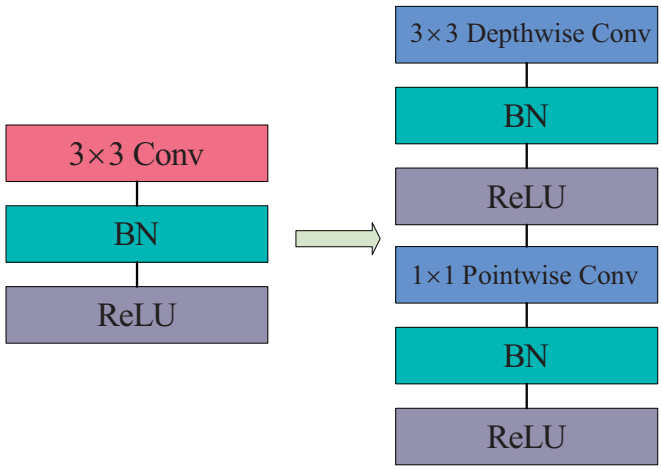


Figure 5. The basic structure of the deep separable convolutional network.

3.1.2. Add Batch Normalization Layer

The SMYOLO model is a fully convolutional network specifically composed of conventional convolution, depth separable convolution, and point convolution. At the same time, the output of each layer of the convolutional network is processed by batch normalization. That is, a batch normalization (BN) layer is added. The proposal of the BN layer solves the gradient disappearance, and gradient explosion in the backpropagation process accelerates the convergence and improves the generalization ability. As shown in the Formula (4), the BN layer uses small batches of normalized convolution features.

$$f = \gamma \times \frac{x - \mu}{\sigma} + \beta \quad (4)$$

Among them, μ and σ represent the mean and standard deviation of the input features. γ and β represent the scale factor and deviation.

3.1.3. Lightweight Activation Function

In YOLOv4, the Mish activation function is one of the innovations of YOLOv4. The Mish activation function is shown in Equation (5). Compared with other commonly used functions such as ReLU and Swish, the Mish activation function is a smooth activation function, which can achieve better accuracy and generalization [19]. Still, Mish activation will generate many calculations, making YOLOv4 unable to complete real-time detection.

Therefore, we use the lightweight activation function Hard-swish, which has similar characteristics with the Mish activation function and reduces the number of model calculations. The comparison between the Hard-swish and Mish activation functions is shown in Figure 6. Hard-swish is shown in Equation (6).

$$Mish(x) = x \times \ tanh(ln(1 + exp(x))) \tag{5}$$

$$Hard - swish[x] = x\frac{ReLU6(x+3)}{6} \tag{6}$$

$$ReLU6 = min(max(0, x), 6) \tag{7}$$

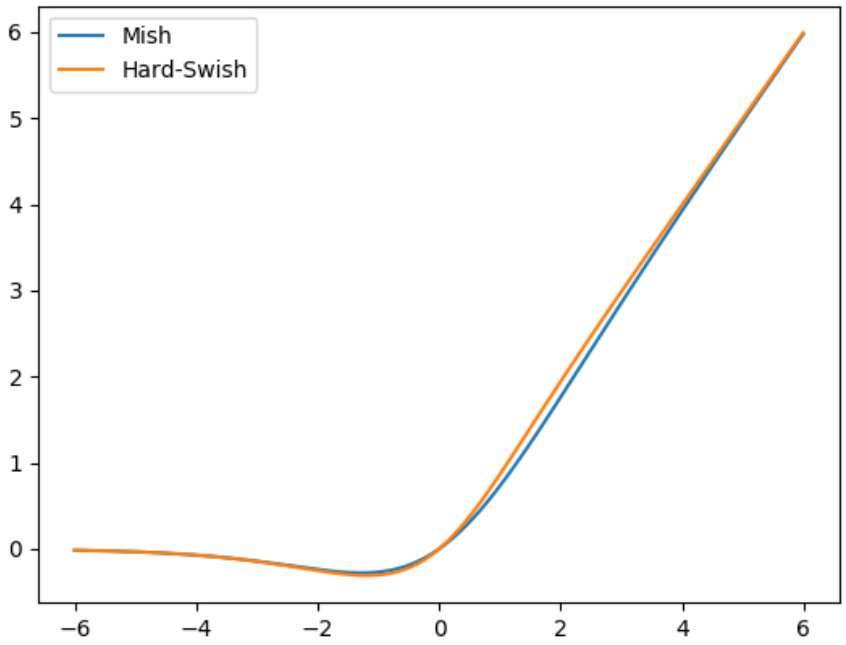


Figure 6. The comparison between the Hard-swish activation function and the Mish activation function.

The overall network architecture of the SMYOLO model is shown in Figure 7, where Convolution represents conventional Convolution, Conv 1 × 1 represents point convolution, and Depthwise Conv represents depth separable Convolution. More profound network architecture can obtain higher accuracy. In contrast, a narrower network architecture limits the complexity of the network. Due to the limited computing power of airborne drone equipment, to limit the complexity of YOLOv4, SMYOLO uses deep separable convolution and point convolution to build the network. In the main convolution operation of the network backbone structure, by adding point convolution, the number of networks increases the depth of the network, thereby reducing the complexity of the model. Compared with other detection networks, SMYOLO has fewer network layers and fewer calculations, enabling it to deploy airborne drone equipment.

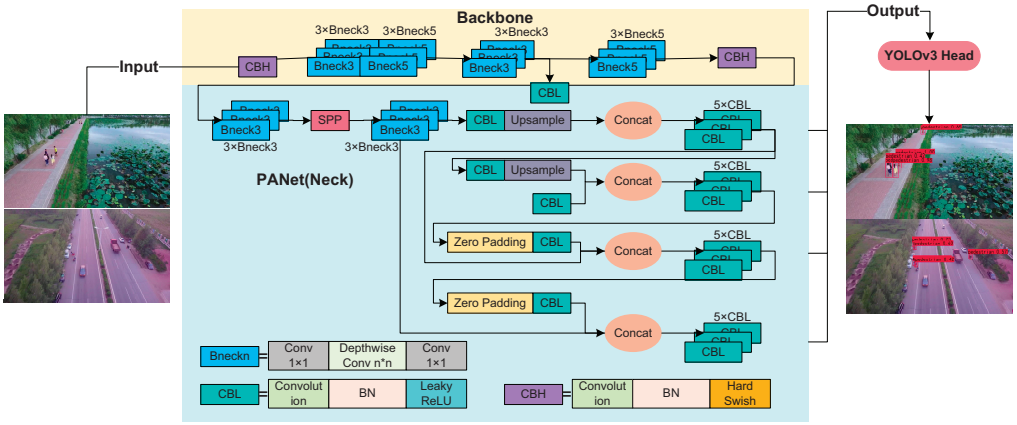


Figure 7. The overall network architecture of the SMYOLO model.

3.2. Improve the Accuracy of SMYOLO through Scale Matching

In low-altitude scenes, because the pedestrian targets captured by aerial photography are small, the practical features of the targets that can be extracted are limited, which is not conducive to detecting pedestrian targets in low-altitude scenes, which leads to low accuracy. In addition, SMYOLO also has a certain degree of accuracy loss while compressing, so this paper proposes to match the pre-training dataset with the detector learning dataset, which will improve the feature representation ability of the detector. SMYOLO can achieve more accurate pedestrian detection in low-altitude scenes. Scale Match matches the pre-training dataset with the detector learning dataset to improve feature representation and improve the detector's performance. The scale matching is essential to make the histogram distribution of the pre-training dataset and the detector learning dataset similar.

First, calculate the scale distribution histogram of the target in the detector learning dataset. Then, calculate the average size s_1 of the label box in any picture in the pre-training dataset, select a bin in the scale histogram of the detector learning dataset, and determine the used bin on the size s_2 of the scale-matched label frame; the scale migration ratio $s2/s1$ is obtained. Finally, scale matching is performed on the pictures in the pre-training dataset according to the scale migration ratio. The calculation formula for scale matching is shown in Equation (8). Among them, $s_0 \in [min(s_1), max(s_1)]$, $min(s_1)$, and $max(s_1)$ represent the minimum and maximum size of the object, respectively. p_{size} represents the general density function, the abscissa of the probability density function is the size of the dataset label frame, and the ordinate is the probability density. After the pre-training dataset and the detector learning dataset are scale-matched, the performance of the SMYOLO target detector is improved.

$$\int_{min(s_1)}^{s_0} p_{size}(s_2; E)ds_1 = \int_{f(min(s_1))}^{f(s_0)} p_{size}(s_2; D_{train})ds_2 \quad (8)$$

This article first processes the VisDrone UAV dataset to extract pictures containing pedestrian targets. Secondly, we match the dataset COCO of the pre-trained model with the pedestrian dataset we extracted to enhance the feature representation ability of SMYOLO and increase the accuracy of the model.

4. Experiments

In this section, a series of experiments will be conducted to verify the performance of the proposed method and compare it with the current mainstream methods. The experimental platform is a PC with a @2.0 GHz CPU and 16G memory. The experiment is mainly implemented on Python 3.7 based on PyTorch and is accelerated by NVIDIA GeForce TITAN XP with 12 GB of memory.

4.1. Performance Evaluation Standard

The experimental data is the publicly available VisDrone2021 DET, collected under various weather and lighting conditions using various drone platforms (drones with different models). The training and validation sets contain 6471 pictures, and the test set contains 1580 images. It should be pointed out that we only extract pedestrian targets in this dataset for training and testing. In the end, we have 5886 images in the training set and validation set and 1197 images in the test set. In order to evaluate the performance of the proposed method, the following evaluation indicators will be used for evaluation.

Frame per second (FPS) measures the speed of object detection, representing the number of video frames that the detector can process per second, and the experiment is to test the FPS of different object detectors on a single GPU.

Intersection over union (IoU) represents the overlap rate between the generated candidate bound and the ground truth bound [26]. The ideal situation is complete overlap. That is, and the ratio is 1.

The mAP is the standard for the detection accuracy of the object detector. It is related to the value of IoU. In the experiment, the value of IoU is set to 0.5, and mAP is the area under the curve of precision and recall. The definition of precision and recall are:

$$Percision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Among them, TP represents the number of detection boxes with $IoU > 0.5$, FP represents the number of detection boxes with $IoU \leq 0.5$, and FN represents the number of undetected.

4.2. Experimental Results and Analysis

4.2.1. Comparison of Different Network Architectures

This experiment compares SMYOLO with three other advanced target detection algorithms and calculates the parameter size and inference speed of each model to verify the effectiveness of SMYOLO. The experimental results are shown in the Table 1. Compared with the original model YOLOv4, SMYOLO has reduced the parameter amount by 81%, the model size has been reduced by 76.90%, and the inference speed has increased by 43.29%. The reduction in parameters and model size indicate that the conventional convolution is replaced. For deep separable convolution and point convolution, the network can be made deep and narrow, and the model can be effectively simplified. The increase in forwarding reasoning time reflects that SMYOLO is more suitable for airborne platforms with poor computing capabilities than YOLOv4. Experimental results show that SMYOLO replaces conventional convolutions with deep separable convolutions and point convolutions, making the network deep and narrow, reducing the amount of calculation, and enabling it to be deployed on airborne platforms, and solves the problem of the high computational load of high-precision target detection models.

Table 1. Comparison of experimental results of different network architectures.

Methods	Parameters (M)	Model Volume (M)	Inference Time (ms)
YOLOv3	62.1	237	29.2
YOLOv4	63.9	245	16.4
YOLOv4-Tiny	8.8	23.1	7.1
SMYOLO	12.2	56.6	9.3

4.2.2. The Effect of Using BN and Multi-Scale Training

This experiment compares the detection results of whether to use the BN layer and multi-scale training. “Y” means used, and “N” means not used. The experimental results

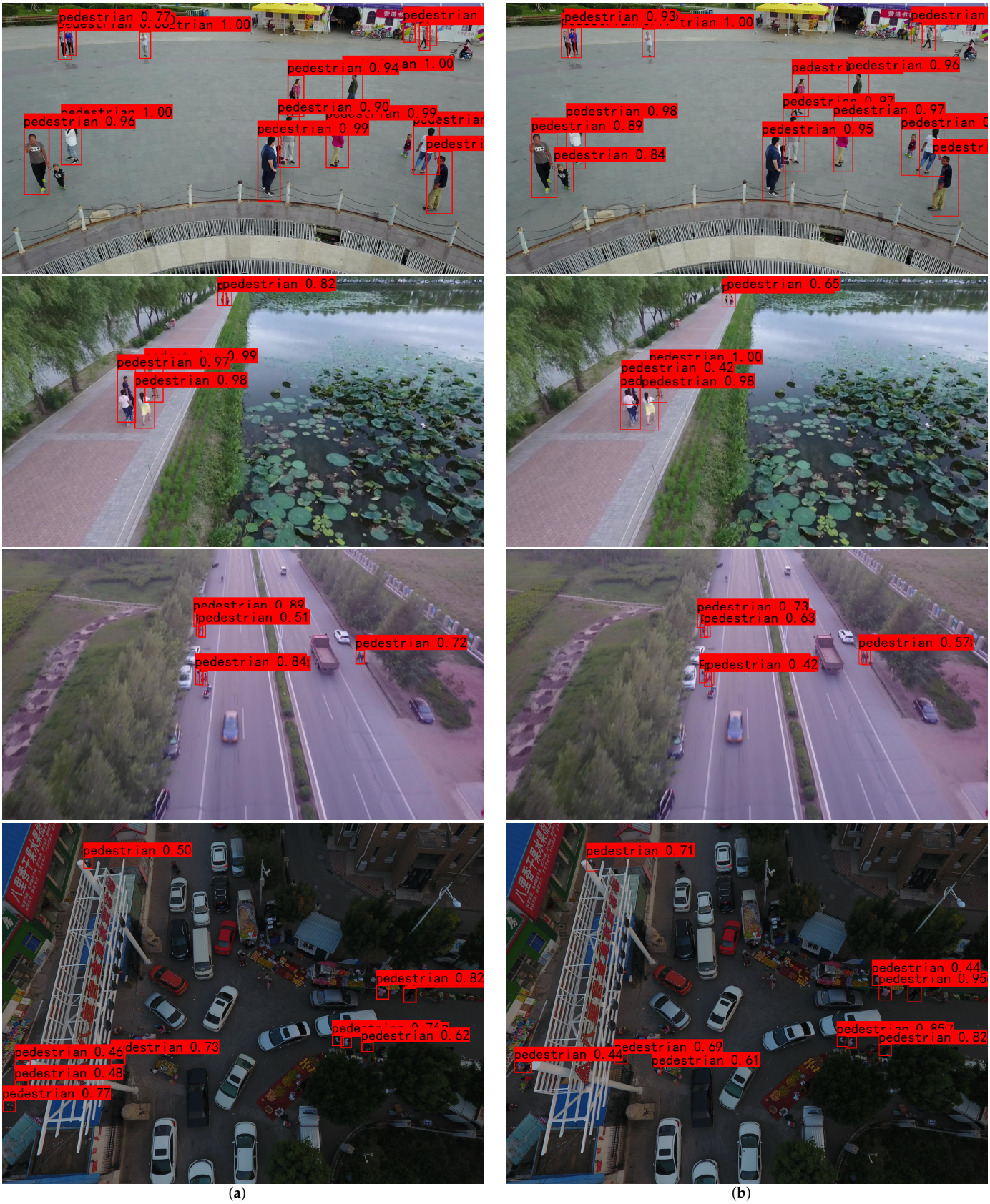
are shown in the Table 2. It can be seen from the table that when SMYOLO only uses the BN layer, the detection accuracy mAP is increased by 3.63%, and the detection accuracy is further increased by 4.28% by using multi-scale training based on the BN layer. Therefore, the SMYOLO network uses the corresponding activation operation of the BN layer for normalization processing and uses multi-scale training in the training process, which improves the accuracy of SMYOLO to a certain extent and enhances its robustness. In addition, in model training, the convergence speed of SMYOLO without the BN layer is very slow, and the time consumption is about four times that of the BN layer. Overall, it is shown that SMYOLO can improve the detection performance by adding multi-scale training based on using the BN layer.

Table 2. Comparison of experimental results of SMYOLO using BN and multi-scale training.

BN Layer	Multi-Scale Training	mAP/%	FPS
Y	Y	58.6	57.72
Y	N	54.32	59.3
N	N	50.69	62.25
N	Y	55.41	58.32

4.2.3. The Effect of the Scale Matching Method

To verify the effect of the proposed scale matching method and show that the proposed method is effective, this experiment compares the detection results of SMYOLO without scale matching and after scale matching. This experiment compares the detection results of whether to use scale matching, and the detection accuracy has been improved. Scale matching can eliminate false positive objects and increase false negative object instances, which improves the performance of the target detector. As shown in Figure 8, in the first example, the missed detection rate is higher before the scale is matched, and after the scale is matched, the missed detection rate is reduced. The overall results show that scale matching can improve detection accuracy. As shown in the figure, in the first example, before the scales are matched, the two children in the picture have small individuals, which increases the difficulty of detection to a certain extent, so there are instances of missed detection. After the scales were matched, the two children were successfully identified, the missed detection rate was reduced, and the missed detection cases were eliminated. In the third example, the target is smaller because the drone is located higher. When pedestrians walk side by side, it will increase the difficulty of detection. Therefore, SMYOLO recognizes the two pedestrians as the same person before the scale is matched. After matching, the missed detection rate is reduced, and the missed detection instances are successfully identified. The overall results show that the scale matching will improve the detection accuracy of SMYOLO.



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

Figure 8. Comparison of the detection results of SMYOLO without scale matching and after scale matching. (**a**) Before scale-match. (**b**) After scale-match.

4.2.4. Compare Experiments with Advanced Detectors

This paper compares the detection accuracy and speed with the five most advanced detectors to verify the performance of the proposed pedestrian detection method, including YOLOv4-Tiny, YOLOv3, YOLOv4, SSD, and SlimYOLOv3. All of the target detectors use the COCO dataset as the pre-training dataset and are trained based on the VisDrone dataset. As shown in Figure 9, compared with the original YOLOv4, the volume of the SMYOLO model is reduced by 76.9%, mAP is reduced by 12.65%, and FPS is increased by 33.25%. After the scale matching, the accuracy is increased by 9.49%. Compared with YOLOv4-tiny, the proposed pedestrian detection algorithm mAP increased by 24.27%. Compared with YOLOv4, the overall FPS increased by 33.33%, achieving the accuracy of mAP and the speed FPS balance, reaching 68.09% and 54.4. As shown in Table 3, the table shows the detailed detection results of different target detectors in the pedestrian dataset extracted by VisDrone. The SSD detector uses a deeper network model, so the accuracy is high. Nevertheless, the complexity of the calculation increases, resulting in real-time performance being inferior. Although YOLOv4 minimally achieves the best real-time performance, it has a poor detection effect on pedestrian targets and low performance. SlimYOLOv3 is an excellent model to improve the YOLO series because this model reduces the complexity of the model through pruning, but the accuracy is not ideal. If deployed on a drone, it cannot achieve the balance of accuracy and delay. In general, the method proposed in this paper improves the detection accuracy while ensuring real-time performance in pedestrian target detection.

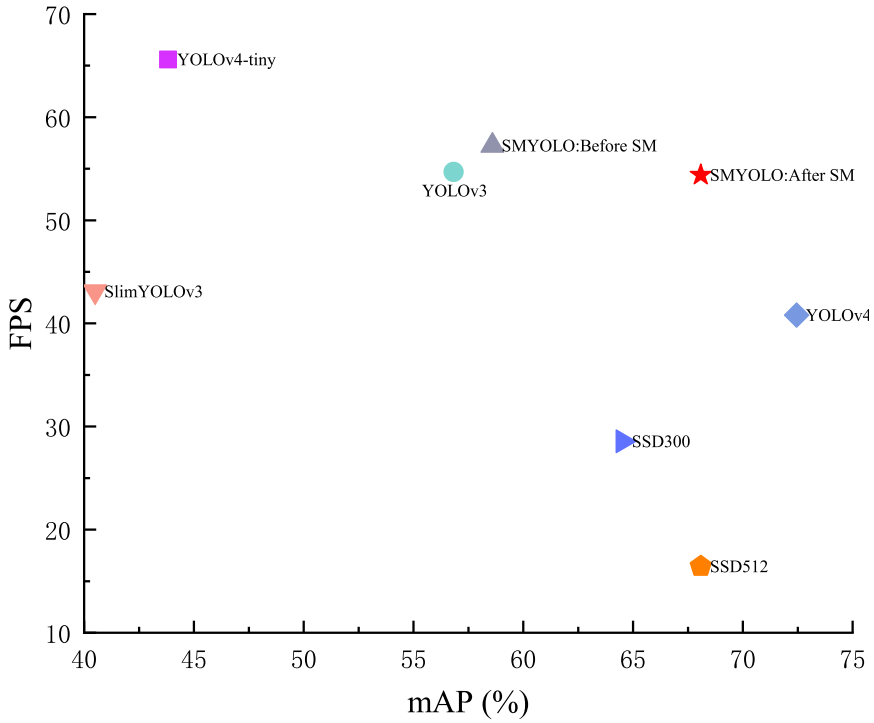


Figure 9. FPS and mAP comparison with mainstream target detection algorithms.

Table 3. Comparison between the proposed method and advanced detection algorithms.

Methods	Input	Model Volume(M)	Precision/%	Recall/%	mAP/%	FPS
SSD300	300 × 300	100	73	57	64.53	28.56
SSD512	512 × 512	104	69	63	68.09	16.44
YOLOv3	416 × 416	237	72.46	49.82	56.83	54.7
YOLOv4	416 × 416	245	76	67	72.45	40.8
YOLOv4 Tiny	416 × 416	23.1	68.37	38.97	43.82	65.6
SlimYOLOv3	416 × 416	79.6	63.4	35.3	40.5	43.1
SMYOLO:Before SM	416 × 416	56.6	72.4	51.68	58.6	57.24
SMYOLO:After SM	416 × 416	56.6	69.4	63.21	68.09	54.4

5. Conclusions

This paper proposes a pedestrian detection algorithm based on deep separable convolution in the UAV scene, which mainly solves the problem of the limited computing power of embedded devices. Secondly, it solves the problem of missed detection of lightweight networks. This algorithm replaces the YOLOv4 conventional convolution with a depth separable convolution, increases the network depth, reduces the network width, and makes the network architecture “narrow and deep", which greatly reduces the number of model parameters and improves the detection efficiency. At the same time, the BN layer and multi-scale training methods are used to improve the model’s accuracy. In addition, we use the scale matching method to solve the problem of feature representation loss caused by scale mismatch, which increases the detection accuracy of SMYOLO to a certain extent. We tested SMYOLO in the same environment. Compared with other detection algorithms, SMYOLO performed more prominently. To ensure the detection accuracy of the model, the detection model has a good detection effect on the target detection on the mobile end of the UAV. We expect to get a faster model while ensuring the same accuracy in the future.

Author Contributions: Conceptualization, W.Z., X.M., and Y.Z.; funding acquisition, W.Z.; investigation, X.M., M.J., and C.Z.; methodology, W.Z. and X.M.; project administration, W.Z.; software, X.M.; supervision, Y.Z. and M.J. and C.Z.; Writing—original draft, X.M.; Writing—review and editing, X.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (Grant No. 61976098), the Technology Development Foundation of Quanzhou City (Grant No. 2020C067), Key Science and Technology Project of Xiamen City (Grant No. 3502ZCQ20201008).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study analyzes publicly available datasets. This datasets can be found here: http://aiskyeye.com/download/ (Accessed on 1 December 2022).

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
SDD	Stanford drone dataset
BN	Batch normalization
SM	Scale match
UAV	Unmanned aerial vehicles
RPN	RegionProposal Network
SE	Squeeze-and-Excitation

References

1. Li, H.; Wu, Z.; Zhang, J. Pedestrian detection based on deep learning model. In Proceedings of the 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, China, 15–17 October 2016; pp. 796–800, https://doi.org/10.1109/CISP-BMEI.2016.7852818.
2. Zhao, Y.; Yuan, Z.; Chen, B. Accurate Pedestrian Detection by Human Pose Regression. *IEEE Trans. Image Process.* **2020**, *29*, 1591–1605, https://doi.org/10.1109/TIP.2019.2942686.
3. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Klette, R. Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. Image Process.* **2017**, *26*, 1992–2004, https://doi.org/10.1109/TIP.2017.2670780.
4. Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; Lecun, Y. Pedestrian Detection with Unsupervised Multi-stage Feature Learning. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Los Alamitos, CA, USA, 2013; pp. 3626–3633, https://doi.org/10.1109/CVPR.2013.465.
5. Chen, S.; Zhang, H.; Lei, Z. Person Re-Identification Based on Attention Mechanism and Context Information Fusion. *Future Internet* **2021**, *13*, 72, https://doi.org/10.3390/fi13030072.
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587, https://doi.org/10.1109/CVPR.2014.81.
7. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. *CoRR* **2015**, abs/1512.02325.
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.
11. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21– July 2017; pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.
12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
14. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Piao, C. UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective. *Sensors* **2020**, *20*, 2238.
15. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale Match for Tiny Person Detection. *arXiv* **2019**, arXiv:1912.10664.
16. Shao, X.; Wei, J.; Guo, D.; Zheng, R.; Zhao, Y. Pedestrian Detection Algorithm based on Improved Faster RCNN. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021.
17. Wu, D.; Lv, S.; Jiang, M.; Song, H. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Comput. Electron. Agric.* **2020**, *178*, 105742, https://doi.org/10.1016/j.compag.2020.105742.
18. Yu, Z.; Shen, Y.; Shen, C. A real-time detection approach for bridge cracks based on YOLOv4-FPM. *Autom. Constr.* **2021**, *122*, 103514, https://doi.org/10.1016/j.autcon.2020.103514.
19. Yang, Y.; Xie, G.; Qu, Y. Real-time Detection of Aircraft Objects in Remote Sensing Images Based on Improved YOLOv4. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021; Volume 5, pp. 1156–1164.
20. Zhao, H.; Zhou, Y.; Zhang, L.; Peng, Y.; Hu, X.; Peng, H.; Cai, X. Mixed YOLOv3-LITE: A Lightweight Real-Time Object Detection Method. *Sensors* **2020**, *20*, 1681, https://doi.org/10.3390/s20071861.
21. Alfasly, S.; Liu, B.; Hu, Y.; Wang, Y.; Li, C.T. Auto-Zooming CNN-Based Framework for Real-Time Pedestrian Detection in Outdoor Surveillance Videos. *IEEE Access* **2019**, *7*, 105816–105826, https://doi.org/10.1109/ACCESS.2019.2931915.
22. Ke, X.; Lin, X.; Qin, L. Lightweight convolutional neural network-based pedestrian detection and re-identification in multiple scenarios. *Mach. Vis. Appl.* **2021**, *32*, 1–23.
23. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. *arXiv* **2018**, arXiv:1803.01534.
24. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2017**, arXiv:1612.03144.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 6–12 September 2014; pp. 346–361.
26. Rahman, Md Atiqur; Wang, Y. Advances in Visual Computing. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. Springer International Publishing, Berlin/Heidelberg, Germany, 10 December 2016, *10072*, 234–244.