electronics

MDPI

Article

Lightweight Pedestrian Detection Based on Feature Multiplexed Residual Network

Mengzhou Sha [1,2], Kai Zeng [1,2,*], Zhimin Tao [3,4], Zhifeng Wang [3] and Quanjun Liu [3]

[1] Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming 650500, China
[2] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[3] Beijing Anlu International Technology Co., Ltd., Mound Stone Road, Beijing 100043, China
[4] Transportation Science and Engineering, Beihang University, Xueyuan Road, Beijing 100191, China
* Correspondence: zengkai@kust.edu.cn; Tel.: +86-180-8002-3451

Abstract: As an important part of autonomous driving intelligence perception, pedestrian detection has high requirements for parameter size, real-time, and model performance. Firstly, a novel multiplexed connection residual block is proposed to construct the lightweight network for improving the ability to extract pedestrian features. Secondly, the lightweight scalable attention module is investigated to expand the local perceptual field of the model based on dilated convolution that can maintain the most important feature channels. Finally, we verify the proposed model on the Caltech pedestrian dataset and BDD 100 K datasets. The results show that the proposed method is superior to existing lightweight pedestrian detection methods in terms of model size and detection performance.

Keywords: autonomous driving; pedestrian detection; multiplexed residual; scalable attention

Citation: Sha, M.; Zeng, K.; Tao, Z.; Wang, Z.; Liu, Q. Lightweight Pedestrian Detection Based on Feature Multiplexed Residual Network. *Electronics* **2023**, *12*, 918. https://doi.org/10.3390/electronics12040918

Academic Editor: Hamid Reza Karimi

Received: 30 November 2022
Revised: 25 January 2023
Accepted: 30 January 2023
Published: 11 February 2023

1. Introduction

Pedestrian detection is a challenging task for autonomous driving [1,2]. With the rapid development of deep learning, some advanced detectors have been constructed with massive weight parameters, such as Mask-RCNN [3], Faster-RCNN [4], SSD [5], and YOLO [6], among others. However, these deep networks cannot guarantee detection efficiency based on insufficient computing resources [7], for example, vehicle or roadside embedded devices. Therefore, researchers developed lightweight networks, which are designed to maintain model accuracy while further reducing the number of model parameters and complexity.

The design of lightweight networks can improve the performance of pedestrian detection for autonomous driving. For the past few years, researchers have made a number of advances in lightweight models [8,9]. For instance, the MobileNet uses deeply separable convolution to cut down parameters [10]. The ResNet adopts residual structures that can effectively avoid the gradient disappearance problem [11]. In order to maintain accuracy and greatly reduce computation cost, ShuffleNet designed group convolution and channel shuffle based on the residual structure to reduce model size [12]. In addition, there are many other advanced lightweight models, such as YOLO v3-tiny [13], YOLO v4-tiny [14], etc.

Although the design of a lightweight model effectively reduces the complexity of the algorithm, their ability to recognize small-scale or scale variation objects, like pedestrians [15], is unsatisfactory. The residual structure is very important for lightweight networks, but the identity shortcut skips the residual blocks to preserve features and consequently might limit the representation power of the network, since the residual connection in the lightweight network is not reusable, resulting in poor recognition of small-scale pedestrians [16]. Attention mechanisms [17] can emphasize the importance of individual features. However, the existing lightweight attention mechanism lacks the perception ability of scale variation [18], resulting in poor multi-scale target recognition [19].

To overcome the weakness of lightweight detectors for recognizing multi-scale and small-scale objects, we propose a novelty feature multiplexed residual network (FMRN). The backbone network is constructed by employing a three-layer multiplexing connection residual block for feature extraction. This lightweight backbone net improves the feature capture ability of small-scale pedestrians. Then, the feature maps are enhanced through the scalable attention mechanism of topology structure. Finally, the targets and classification are obtained after a full connection layer and regression. The contributions of this paper are as follows:

(1) We propose a novel multiplexer residual (MR) method to build the feature extraction network. The multiplexed connection residual structure retains the characteristic information of the previous layer and passes the useful characteristic information to the output of the next layer. The MR improves the information transmission ability of the traditional methods, which is more conducive to the lightweight model to capture small-scale pedestrian features;
(2) A lightweight scalable attention module (SA) is investigated to expand the respective field of the detection model. The branch structure of the SA module is selected to synchronize the feature dimensions, and the dilated convolution is introduced to expand the local respective field of the model. The SA module can eliminate redundant channel information, which can further improve the adaptation ability of the model to deal with the issue of pedestrian scale variation;
(3) Experiments show that our proposed FMRN model is superior to existing lightweight pedestrian detection methods. Our model can reach 66.4% detection accuracy in the Caltech dataset, model size (17.6 Mb), detection speed (FPS 124), and excellent detection performance.

The rest of this paper can be summarized as follows: Section 2 summarizes related studies; Section 3 describes the design details of the model and the implementation details of each innovation point in detail. A large number of comparative experiments which verify the advanced nature of this method are demonstrated in Section 4. Section 5 summarizes the conclusion of this paper and the directions for further research in this field.

2. Related Studies

With the popularity and development of autonomous driving, intelligent sensing pedestrian detection has gradually become a research hotspot. Researchers have focused on the balance between lightweight models and detection performance. Many excellent network reduction algorithms have been proposed, such as two-stage detection models [20] for generating candidate frames and end-to-end single-stage detection models [21], but the two-stage detection models are slow in inference. The single-stage detection model is less effective at detecting difficult samples, such as small-scale pedestrians; hence, how to ensure the real-time performance of the algorithm and enhance network detection [22] has become a hot topic of research.

2.1. Single-Stage Detection Model

In recent lightweight research, there are many advanced one-stage detectors. Yi et al. [23] improved the YOLO v3-tiny backbone network by adding three convolutional layers and introducing a 1×1 convolutional kernel to reduce the complexity of the algorithm. However, they only used three convolutional layers to the backbone network, which led to some improvement in the detection performance of the algorithm, but the false detection of pedestrian misses was still high. The basic unit of ENet [24] is residual structure, and the detection effect is better than that of the VGG network [25]. This is due to the network structure design of ENet encoding and decoding, which adds information compilation at different network layers. To achieve real-time pedestrian detection speed without reducing detection accuracy, Murthy et al. [26] proposed an optimized MobileNet combined with an SSD network and added contextual information using a connected feature fusion module. MobileNet uses separable convolution to build a feature extraction network, which greatly

reduces convolutional computation [27]. In the face of the complex scene changes of autonomous driving, the information interaction capability of MobileNet is poor, and the transfer of deep and shallow information during model feature extraction is not considered, resulting in a small model size but poor pedestrian detection. Shao et al. [28] took PeleeNet as the backbone and further integrated multi-scale features and spatial concerns to enhance the characteristics of small objects, such as people.

At present, the residual structure is widely used for designing the network in the field of lightweight pedestrian detection. One of the reasons that makes ResNet exceptionally popular is the simple design strategy, which introduces only one identity shortcut. However, the identity shortcut might limit the representation power of the network [29]. Moreover, it causes the collapsing domain problem [30], which weakens the network's ability to detect small-scale pedestrians.

2.2. Lightweight Attention Mechanism

Other researchers have improved the detection performance of lightweight models by designing attention mechanisms [31]. Wang et al. [32] proposed a multi-scale pedestrian detector APNB + ASFF based on a self-attention mechanism and adaptive spatial feature fusion. They used an attention mechanism to solve the problem of poor small-scale pedestrian detection, but the attention mechanism module also involves a large number of parameters. Current state-of-the-art attention mechanisms in the field of pedestrian detection include PPM [33] and RFB [34], among others. Yu et al. [33] used CNN as the backbone for feature extraction and used the attention mechanism PPM to capture important details in the images, and multi-scale features were effectively fused to gain cross-channel attention. Zeng et al. [34] replaced the convolutional layers with RFB structures in the two output feature layers of the SSD detection network. The improved algorithm showed a significant improvement in the detection results of the KITTI dataset. Designing more effective lightweight attention mechanisms is gradually becoming a research hotspot.

The attention mechanism can emphasize the importance of features and improve the detection effect of the lightweight model [35]. Most existing approaches focus on developing more complex attention modules for better performance, which inevitably increases the complexity of the model. However, the existing lightweight attention mechanism lacks information perception of different dimensions and cannot capture pedestrian characteristics at different scales [36]; thus, it is difficult to effectively detect pedestrians at multiple scales.

3. Methods

In this section, we study the network structure of FMRN in detail. It is composed of two parts: a multiplexing connection residual structure and a scalable attention mechanism. We introduce the design idea of a multiplexing connection residual structure in Section 3.2. The scalable attention mechanism is introduced in this subsection. Then, the structure of the loss function is described in the Section 3.4.

3.1. Overall Networks

The network structure of our model is shown in Figure 1. The overall network consists of a convolutional layer, a pooling layer, a reuse connection residual structure, and a scalable attention mechanism. The initial input size of the image is $416 \times 416 \times 3$. The size of the feature map at the convolution and pooling layer is $416 \times 416 \times 16$. The feature dimensions increase after multiplexing. It can enhance the ability of the residual structure for retaining its characteristics. The scalable attention mechanism module is cascaded behind the feature layer to expand the local perceptual field by using different expansion rates of convolutions and adapting the scale variation of pedestrians. It introduces an attention mechanism to filter pedestrian features that are subject to background and occlusion to optimize pedestrian detail features. The structure of our proposed network model is shown in Figure 1.

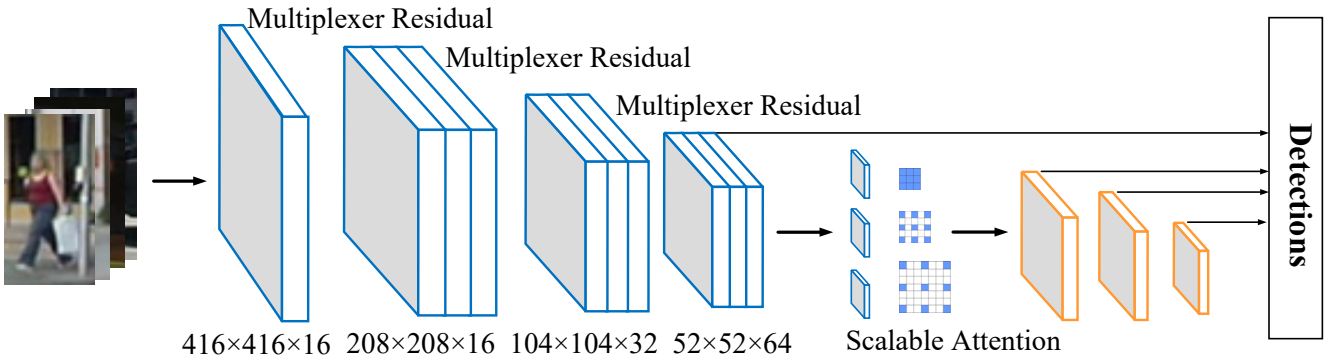


Figure 1. The network structure of FMRN.

The feature extraction network is built based on the multiplexed connection residual blocks to reduce the size of the network model. The FMRN network parameters are shown in Table 1.

Table 1. FMRN network parameters.

Network Layer	Convolution Kernel Number	Convolution Kernel Size	Output Size
Conv		$3 \times 3/1$	416×416
maxpooling		$2 \times 2/1$	208×208
Multiplexer Residual (1)	16	$\begin{bmatrix} 16 & 1 \times 1 \\ 16 & 3 \times 3 \end{bmatrix}$	208×208
Conv		$1 \times 1/1$	208×208
maxpooling		$2 \times 2/2$	104×104
Multiplexer Residual (2)	32	$\begin{bmatrix} 16 & 1 \times 1 \\ 16 & 3 \times 3 \end{bmatrix}$	104×104
Conv		$1 \times 1/1$	104×104
maxpooling		$2 \times 2/2$	52×52
Multiplexer Residual (3)	64	$\begin{bmatrix} 16 & 1 \times 1 \\ 16 & 3 \times 3 \end{bmatrix}$	52×52
Conv		$1 \times 1/1$	52×52
maxpooling	128	$2 \times 2/2$	26×26

3.2. Multiplexing Connection Residuals

ResNet [37] first proposed the idea of residual structure and jump connection, which change the output of a certain layer into a linear superposition of the input and a nonlinear transformation of the input. This structure not only solves the problems brought to the network by the deepening of the number of convolutional layers but also makes the information transfer more effective. Take layer i as an example, and the input of layer $i + 1$ as:

$$X_{i+1} = X_i + F(x_1, W_1) \tag{1}$$

where, X_{i+1} represents the output data, X_i represents the input data, W_1 is the weight of the neurons, and $F(.)$ represents the result of the input data through the residual structure.

The conventional residual structure uses a ReLU activation function with a derivative of 0 when z is less than 0. Neuron death may occur during gradient descent, whereas Leaky ReLU still has parameters in negative coordinates that prevent the gradient problem that occurs when the network is backward. Figure 2 shows two mathematical models of activation functions.

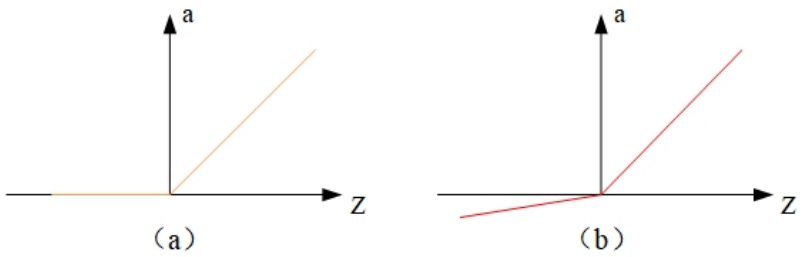


Figure 2. (**a**) ReLU activation functions; (**b**) Leaky ReLU activation functions.

Figure 3a shows a schematic diagram of the residual structure and the multiplexed connection residual structure. Traditional convolutional or pooling layers are prone to information loss when transmitting the information. The network model lacks the ability to generalize. The whole network usually learns only the feature difference part, simplifying the difficulty and steps of learning. As shown in Figure 3b, we propose a multiplexed connected residual block of the same latitude based on the residual bottleneck structure. It introduces 1×1 convolution to reduce the computation while maximizing the information flow between all layers in the network. The interaction between deep and shallow information is enhanced when extracting pedestrian features. Thus, we can build a feature extraction network using the multiplexed connected residual structure.

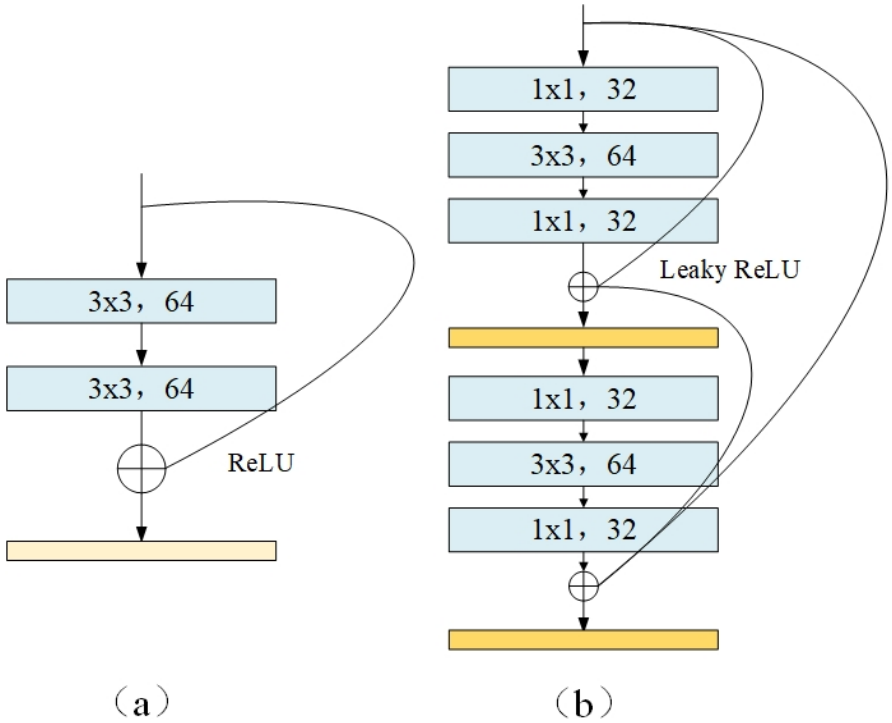


Figure 3. (**a**) Residual structure of ResNet; (**b**) Multiplexing connection residual structure.

3.3. Scalable Attention Mechanism

In this section, we study the scalable attention mechanism that employs dilated convolution to obtain a larger perceptual field. It merges the input feature channels through data filtering to extract information that is more valuable to the classification of the network model. The scalable attention mechanism is shown in Figure 4.

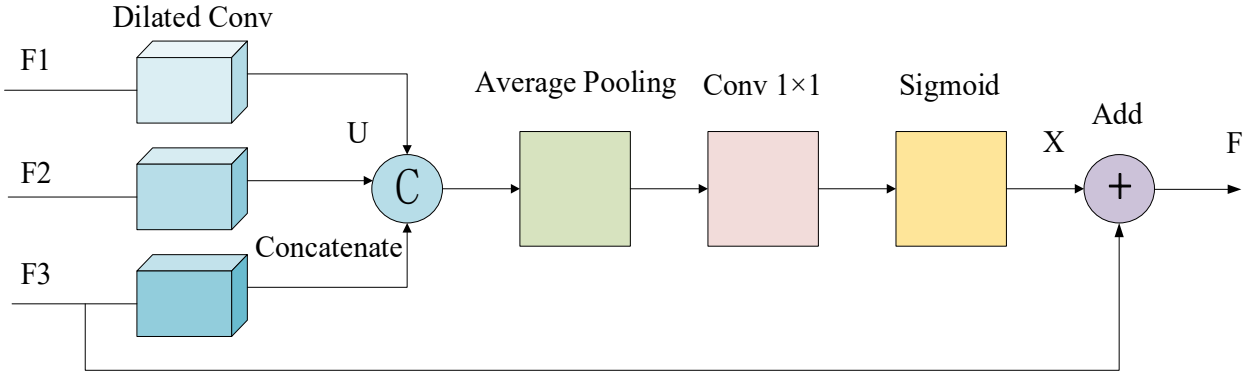


Figure 4. Scalable attention mechanism.

The concatenation operator is given by Equation (2). The vectors F_1, F_2 and F_3 are concatenate connected after dilated convolution of the multi-branch structure.

$$U = Concat\{DilateConv(F_1), DilateConv(F_2), DilateConv(F_3)\} \tag{2}$$

Firstly, the number of channels in the input feature map is carried out by average pooling to shrink the spatial dimension of U. The output of each channel is a scalar, and the calculation formula is given by Equation (3):

$$Z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{3}$$

where U_c represents the input feature channel and Z_c indicates the spatial dimension of global average pooling compression.

Secondly, the sigmoid function is used to obtain the weight of each channel. Non-linear interactions between channels can learn non-exclusive relationships and obtain the importance ratio of each channel. The calculation formula is given by Equation (4):

$$s = F_{ex}(z, W) = \delta(g(z, W)) = \delta(W_2 \delta(W_1, z)) \tag{4}$$

where $W_1 \in R^{\frac{C}{r} * C}$ represents dimension reduction to $\frac{C}{r}$, $W_2 \in R^{C * \frac{c}{r}}$ indicates dimension increase to dimension vector C, and δ represents the ReLU function, which indicates a full connection layer.

After the 1×1 convolution layer and sigmoid activation layer, the attention coefficient between 0 and 1 is obtained. Non-linear interactions between channels can learn non-exclusive relations and obtain the importance ratio of each channel. The calculation formula is given by Equation (5):

$$X = F_{scale}(u_c, s_c) = s_c u_c \tag{5}$$

where F_{scale} represents multiplication, X represents the output of a new feature graph, u_c represents features and s_c represents scalars.

Finally, we add the coefficient to achieve data filtering and extract more valuable information for network model classification. The calculation formula is given by Equation (6):

$$F = X + F_3 \tag{6}$$

3.4. Loss Function

In autonomous driving scenes, the physical size of pedestrians is small, and the detection network is easily disturbed by scale variation. Therefore, the FMRN loss function is obtained from the sum of three parts, which are the center coordinate and width-height coordinate error L_{pos} of the pedestrian object, the confidence error L_{obj}, and the classification error L_{cls}, respectively. The specific calculation formula is as follows:

$$L_{pos} = \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{obj} \left\{ [(x_i - \hat{x}_i{}^j)^2 - (y_i - \hat{y}_i{}^j)^2] + \left[\left(\sqrt{\omega_i^j} - \sqrt{\hat{\omega}_i^j} \right)^2 + \left(\sqrt{h_i^j} - \sqrt{\hat{h}_i^j} \right)^2 \right] \right\} \quad (7)$$

$$L_{obj} = \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{noobj} (c_i - \hat{c}_i)^2 + \lambda_{obj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{obj} (c_i - \hat{c}_i)^2 \quad (8)$$

$$L_{cls} = \sum_{i=0}^{s^2} I_{i,j}^{obj} \sum_{c \epsilon class} ([\hat{P}_i^j \log(P_i^j) + (1 - \hat{P}_i^j) \log(1 - P_i^j)]) \quad (9)$$

where λ_{coord} is used to coordinate the different sizes of the rectangle box contributions to the error function and is not a consistent set of coordination coefficients; s^2 is the size of the feature graph; B is the number of prior frames; x_i, y_i are the horizontal and vertical coordinates of the center point; ω_i,h_i is the width and height of the prediction box, respectively; λ_{noobj} represents the weight of the confidence error in the loss function when the prediction box does not predict the target; $I_{i,j}^{obj}$ represents the first j anchor frame in the first i grid, which is 1 if there is a pedestrian target and 0 otherwise; similarly, $I_{i,j}^{noobj}$ represents the fact that there is no pedestrian in the first j anchor frame in the grid; c_i represents the probability score that the prediction box contains the target object; $\hat{c}_i$ represents the true value; λ_{obj} represents the weight of the confidence error in the loss function when the target is predicted in the prediction box; P_i^j represents the probability that the (i, j) prediction box belongs to category c; and $\hat{P_i^j}$ indicates the true value of the category to which the tag box belongs.

4. Experiments and Analysis of Results

We first introduce two kinds of pedestrian detection datasets and the purpose of pretreatment, i.e., Caltech [38] and BDD 100 K [39]. Experimental equipment and an evaluation metric are presented. Then, the implementation details of our FMRN model are described. To demonstrate the effectiveness of our multiplexing connection residual and scalable attention module, we make some ablative studies. Finally, our model is compared with a state-of-the-art lightweight pedestrian detection network.

4.1. Experimental Dataset and Experimental Parameters

Caltech Pedestrian dataset is the largest Pedestrian dataset in the field of autonomous driving. Video data are collected by vehicle cameras throughout the whole process, including a total of 10 h of 30 Hz video of 640 × 480 pixels, mainly in rural streets. In order to eliminate the influence of inter-frame information of video data on detection results [34], one image out of every fourteen datasets was selected to retain the original format in the data preprocessing part, and a total of 4389 training sets and 4340 test sets were obtained. Figure 5 shows an example of the training set and test set sections of the Caltech pedestrian dataset.

The BDD 100 K dataset released by Berkeley University is a challenging dataset of traffic scenes, collected from across the United States. The dataset covers driving images at different times of day, such as early morning, midday, evening, and night, and also contains many complex weather scenarios, such as rainy, cloudy, and snowy days. In this paper, we select images from the BDD 100 K dataset where only pedestrian targets are present and construct the sub-dataset BDD 100 K-Person by taking only the 5th frame image. A total of 4420 training sets and 3220 test sets were obtained, and this dataset was used as an experimental supplement to the Caltech pedestrian dataset. Figure 6 shows a partial example of the BDD 100 K dataset.

Figure 5. Example of Caltech Pedestrian dataset.

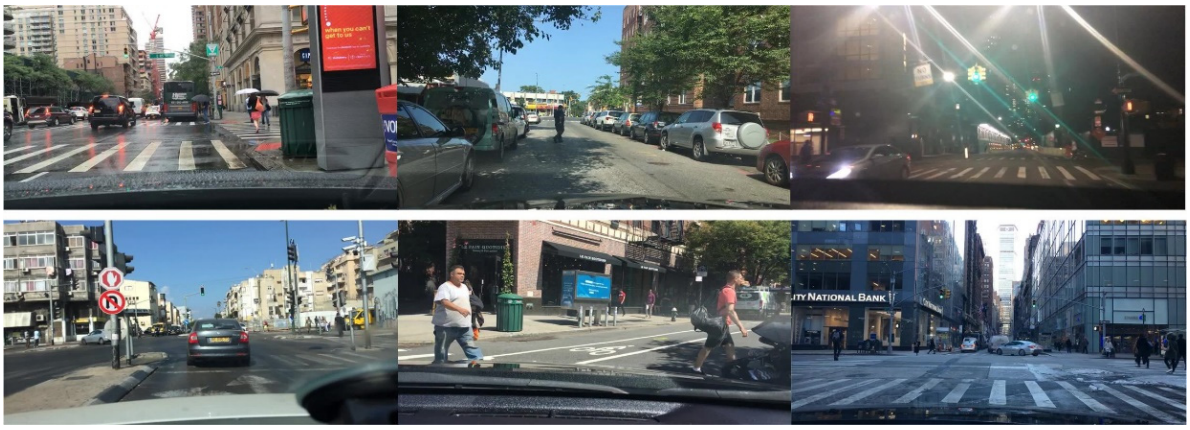


Figure 6. Example of BDD 100 K-Person dataset.

All the experiments in this paper used Ubuntu 16.04 as the main system; the workstation processor model was NVIDIA GeForce RTX 2060; and the memory was 16 G. The deep learning framework adopts the framework and image processing library that are commonly used in autonomous vehicle engineering. The experimental facilities and parameter configurations are shown in Table 2.

Table 2. Experimental environmental parameters.

Item	Parameter
CPU	Intel Core i5-9400F 2.9 GHz x6
GPU	NVIDIA GeForce RTX 2060
Operating system	Ubuntu 16.04 LTS
Memory	16 GB
Deep learning framework version	Pytorch 1.8
Development languages	Python 3.6

In this experiment, the initial network input size was set as 416 × 416, momentum was set as 0.9, batch size of each round was set as 8, the learning rate was 0.001, and weight decay was 0.05, ensuring the fairness of the experiment. All the comparison experiments adopted the same parameter settings. Figure 7 shows the loss curve of network training.

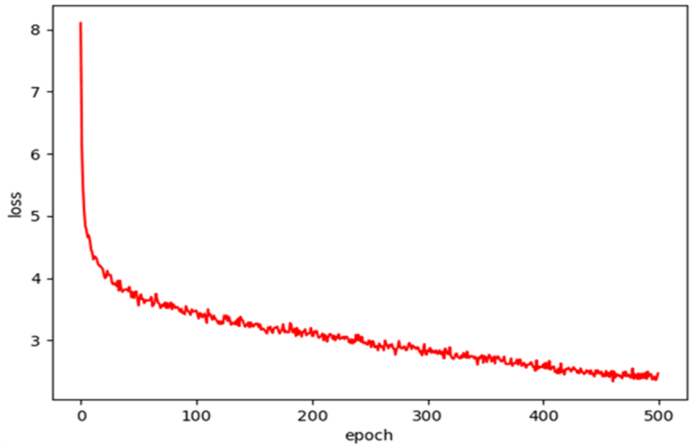


Figure 7. Training loss curve.

4.2. Evaluation Indicators

The evaluation indexes of object detection algorithms mainly include detection accuracy and detection speed. Average Recall (AR) is the ratio of detected recognition frames to real frames of a certain category. The mathematical relation of the missed rate is the higher the Recall rate, the better, and the lower the missed rate, the better. Average Precision (AP) is especially suitable for the algorithm that simultaneously predicts the position and category of objects. It represents the area value of the P–R curve at different IOU values (IOU is 50% in this paper). The larger the AP value, the higher the average accuracy of the model. Another important evaluation index of pedestrian detection algorithm is speed. FPS (Frame Per Second), which is defined as the number of images that can be processed per second, is used to evaluate the speed of pedestrian detection.

$$AR = \frac{TP}{TP + FN} = \frac{TP}{all\ groundtruth} \tag{10}$$

$$AP = \sum_{k=1}^{n} (r_{k+1} - r_k) \times p_k \tag{11}$$

where TP represents correctly identified pedestrians, FN represents positive samples incorrectly identified as negative samples, and all groundtruth represents all targets to be identified.

4.3. Ablation Experiments

The baseline model of this paper is YOLO v3-tiny network. The measurement unit of the model parameter is megabytes (Mb), and the measurement unit of detection speed is FPS, i.e., the number of frames transmitted per second. Based on the Caltech Pedestrian detection dataset, an ablation experiment was conducted for the innovation points in this chapter. It can be seen from Table 3 that:

(1) After using the multiplexed connection residual structure, the feature extraction network can accurately extract pedestrian targets in the complex traffic background; the missed detection rate of the pedestrian is reduced by 4.1%; the average detection accuracy is increased by 2.1%; the number of parameters of the model is reduced by about 50% compared with that of the baseline network; and the model parameter size is only 17.2 Mb. The inference speed of the algorithm is faster than that of the baseline network, and it satisfies the real-time requirement well;

(2) The SA module proposed in this paper does not impose an additional burden on the detection network; the number of model parameters does not increase significantly; the missed detection rate is further reduced by 1%; and the average detection accuracy is increased by 1.3%. The ablation experiment of the Caltech pedestrian data set

showed that the FMRN recall rate, detection accuracy, and model size of this model reached 64.5%, 66.4%, and 17.6 MB, respectively, which were better than those of the baseline network. The missed detection rate was reduced by 5.1% and the average detection accuracy was increased by 3.4%.

Table 3. Ablation experiments.

Experiment Number	Multiplexed Residual	Scalable Attention	AR_{50} (%)	AP_{50} (%)	Model Size (Mb)	FPS
1			59.4	63	34.8	120
2	√		63.5	65.1	17.2	128
3		√	62.6	65.7	35	117
4	√	√	64.5	66.4	17.6	124

The attentional mechanism can effectively enlarge the local receptive field of the model and remove the redundant information of feature channels. Attention mechanisms commonly used in pedestrian detection are mainly divided into channel domain and spatial domain. In recent years, the channel domain and spatial domain have evolved into various morphed attention mechanisms.

To verify the effectiveness of the scalable attention mechanism, this chapter selected the attention mechanism widely used in pedestrian detection algorithms and the scalable attention mechanism (SA) proposed in this chapter for comparative experiments, including Squeeze and Excitation (SE) [40], Convolutional Block Attention Module (CBAM) [41], Pyramid Pooling Module (PPM) [33], and Receptive Field Block (RFB) [34], where PPM and RFB are the latest lightweight attention mechanisms in the field of pedestrian detection and improve the detection effect significantly.

The experimental results of the Caltech pedestrian dataset are shown in Table 4. Compared to the baseline network, our proposed scalable attention mechanism reduces the detection miss rate by 5% and increases the average detection accuracy by 2.7% in the Caltech pedestrian dataset. Compared to the latest lightweight attention mechanisms PPM and RFB applied in pedestrian detection, the scalable attention mechanism (SA) of this chapter is more effective, with a 1.4% and 1.8% reduction in missed detection rate, 0.7% and 1.5% improvement in average detection accuracy over PPM and RFB, respectively, as well as fewer model parameters and faster inference.

Table 4. Comparative experiments on attention mechanisms for Caltech detection datasets.

Model	Dataset	Attention Model	AR_{50} (%)	AP_{50} (%)	Model Size (Mb)	FPS
Baseline	Caltech Pedestrian	-	59.4	63	34.8	120
		+SE	62.1	63.4	35.3	106
		+CBAM	63.2	64.7	35.3	105
		+PPM	63	65	72.6	78
		+RFB	62.6	64.2	55.2	90
		+SA	64.4	65.7	35	117

To further validate the effectiveness of the scalable attention mechanism, the scalable attention mechanism proposed in this chapter was also subjected to comparative experiments on the BDD 100 K dataset under the same parameter configuration. As can be seen from Table 5, the scalable attention mechanism reduces the missed detection rate by 4.2% and increases the average detection accuracy by 3.4% in the BDD 100 K dataset, expanding the model's respective field and capturing small-scale pedestrian features more effectively with essentially the same number of model parameters and inference speed. The scalable attention mechanism captures more detailed information, expands the model's local field of perception without creating complex computational problems, and effectively improves the performance of the pedestrian detection network.

Table 5. Comparative experiments on attention mechanisms for BDD 100 K-Person datasets.

Model	Dataset	Attention Model	AR_{50} (%)	AP_{50} (%)	Model Size (Mb)	FPS
Baseline	BDD 100 K-Person	-	28.2	28.4	34.8	106
		+SE	28.3	28.4	35.3	90
		+CBAM	31.3	30.2	35.3	87
		+PPM	31.6	30.9	72.6	70
		+RFB	32.1	29.5	55.2	78
		+SA	32.5	31.8	35	98

4.4. Comparative Experiment of Lightweight Model

This chapter compares the proposed model with current state-of-the-art lightweight pedestrian detection networks on the Caltech Pedestrian Dataset. YOLOX [42] is a high-performance Anchor free detector, which adds decoupling head, Anchor free, and advanced label allocation strategy to the network. Xception + SSD [43] represents an improved version of ShuffleNet [12] and the SSD algorithm. The main idea is to optimize the backbone network of the SSD algorithm by using an inception structure, thus achieving higher detection accuracy and less computation. The main idea of MobileNet + SSD [26] is to use MobileNet to optimize SSD network parameters. MobileNet mainly uses separable convolutional design features to extract the network and reduce the complexity of the model. APNB + ASFF [32] is a multi-scale pedestrian detector based on a self-attention mechanism and adaptive spatial feature fusion, which uses a lightweight attention mechanism to solve the problem of poor detection effect of the small-scale pedestrian.

As can be seen from Table 6, the detection recall of this chapter's method reaches 64.5% and the average detection accuracy reaches 66.4%, both of which are better than the current mainstream lightweight pedestrian detection networks. Compared with the latest pedestrian detection methods ResNet10 and APNB + ASFF, our model has a 1.0% lower missed detection rate, 1.2% and 1.4% higher average detection accuracy, respectively, and has fewer model parameters and faster detection speed. The experimental results show that the method in this chapter is suitable for autonomous pedestrian detection because it can improve small-scale and scale variation pedestrian detection while effectively reducing the number of model parameters.

Table 6. Caltech Pedestrian dataset lightweight model comparison experiment.

Dataset	Binary Model	AR_{50} (%)	AP_{50} (%)	Model Size (Mb)	FPS
Caltech Pedestrian	ENet	62.3	62.4	26.7	52
	YOLOX	63	64.8	54.2	102
	SSD(VGG)	-	57.8	110.2	37
	Xception-SSD	52	61	56.8	74
	MoblieNet-SSD	63.1	62.6	43.6	85
	ResNet10	63.5	65.2	24.1	88
	APNB+ASFF	63.5	65	61.2	64
	Our model	64.5	66.4	17.6	124

The application of the traditional convolution leads to losses of start-up formation and implicitly to the loss of information in the transmission process. We design a multiplexed connected residual structure, which, using convolution 1×1 and residual structure, not only reduces the burden of computer operation but also maximizes the flow of information between all layers in the network. Therefore, in comparison to other experiments, our FPS has a certain advantage.

The BDD 100 K dataset is a complex and variable scene, containing a variety of challenging images with low light and strong light interference. We conducted the same comparative experiments on the BDD 100 K dataset, and, as shown in Table 7, our proposed model FMRN can achieve 38.9% detection recall and 38.7% detection accuracy, respectively,

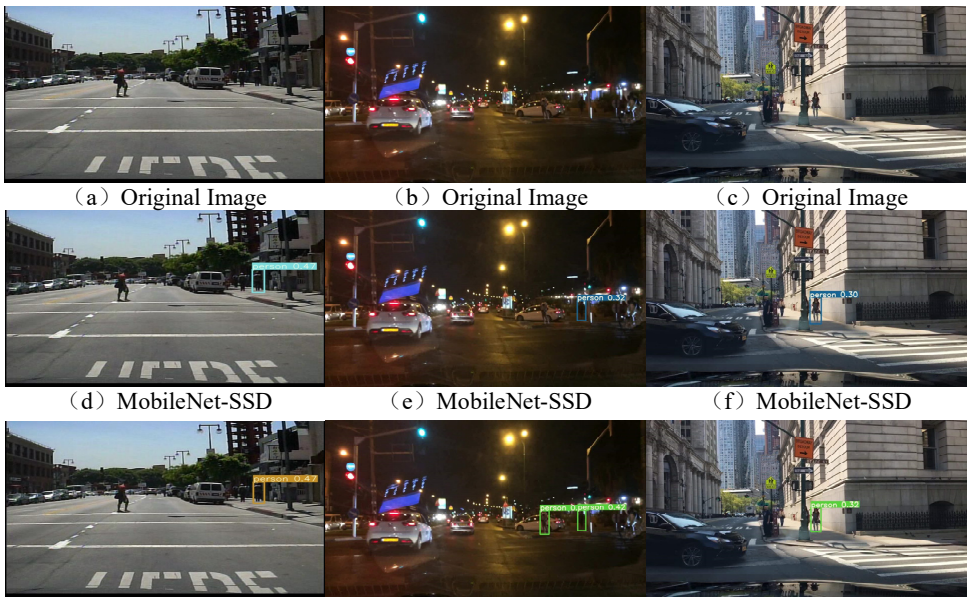
which are both better than current lightweight pedestrian detection methods. Compared with the latest pedestrian detection method PeleeNet [44], the method in this chapter has a 1.4% reduction in missed detection rate and a 0.7% improvement in average detection accuracy, as well as a significant reduction in model size and a substantial improvement in detection speed. The experimental results show that the FMRN model has a simple structure and is easily portable on GPU devices with low computational performance.

Table 7. BDD 100 K-Person dataset lightweight model comparison experiment.

Dataset	Binary Model	AR_{50} (%)	AP_{50} (%)	Model Size (Mb)	FPS
BDD 100 K- Person	ENet	32.3	30.5	26.7	43
	YOLOX	33.2	34.2	54.2	88
	SSD(VGG16)	-	33.4	110.2	30
	Xception-SSD	35.6	37.1	56.8	67
	MoblieNet-SSD	36.2	38.5	43.6	78
	ResNet10	32.1	30.6	24.1	72
	PeleeNet	37.5	38	24.6	89
	Our model	38.9	38.7	17.6	101

4.5. Detection Visualization

We give three representative pedestrian detection networks for visualization; the detection results are shown in Figure 8. The interaction of deep and shallow information in the image is facilitated by the reduced loss of information due to the multiplexed connected residual structure in the convolutional pooling layer when extracting pedestrian features; hence, the multiplexed connected residual enhances the network's ability to capture small-scale pedestrian features and semantic information. In addition to pedestrian targets facing complex scale variation and the scalable attention to design dilated convolution modules with different branching structures, the attention mechanism can adapt to the complex scale variations of pedestrians and has good detection results for multi-scale pedestrians.



（a）Original Image （b）Original Image （c）Original Image

（d）MobileNet-SSD （e）MobileNet-SSD （f）MobileNet-SSD

（g）ResNet10 （h）ResNet10 （i）ResNet10

Figure 8. *Cont.*

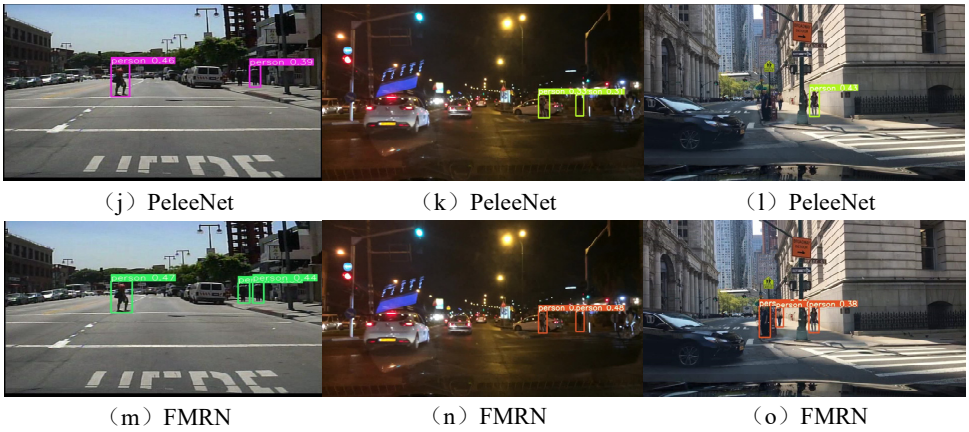


Figure 8. Visual comparison with previous state-of-the-art methods on Caltech and BDD 100 K-Person dataset.

5. Conclusions

In Sections 4.2–4.4, we not only conduct ablation experiments to prove the effectiveness of each module, but also compare experiments in two datasets to show that our model has a small number of parameters, fast detection speed, and good detection effect. The main reasons are as follows.

The multiplexed connection residual structure (MR) retains the characteristic information of the previous layer and passes the useful characteristic information to the output of the next layer. The MR improves the information transmission ability of the traditional methods, which is more conducive to the lightweight model to capture small-scale pedestrian features.

A lightweight scalable attention module (SA) is investigated to expand the respective field of the detection model. The branch structure of the SA module is selected to synchronize the feature dimensions, and the dilated convolution is introduced to expand the local respective field of the model. The SA module can eliminate the redundant channel information, which can further improve the adaptation ability of the model to deal with the issue of pedestrian scale variation.

Pedestrian detection is of profound importance to autonomous driving. This paper proposes a lightweight pedestrian detection method based on a multiplexed connection residual network. Firstly, a multiplexed connection residual structure is designed based on the residual structure idea, and a new feature extraction network is built on YOLO v3-tiny network using this structure. Then, a scalable attention mechanism module is proposed to expand the model's receptive field and enhance the feature extraction capability of the detection network for small-scale pedestrians. Experimental results show that the proposed method is lighter than YOLO v3-tiny, with only 17.6 MB of parameters. Validation experiments on the Caltech pedestrian dataset and BDD 100 K pedestrian dataset prove that the proposed method can reduce the number of parameters in the network model and improve the detection performance for pedestrians, especially for small-scale pedestrians.

This research can bring different research ideas to the application of lightweight models, pedestrian detection, and other computer vision fields, to help develop more lightweight models to bring better detection results. In the future, we will focus on the integration of lightweight detection models and multi-mode fusion technology, explore the joint detection of infrared images or radar sensor information, and strengthen the robustness of the pedestrian detection network in cases of bad weather.

Author Contributions: Conceptualization, M.S.; methodology, M.S.; software, M.S.; validation, M.S.; formal analysis, K.Z.; investigation, K.Z.; resources, Z.T., Z.W. and Q.L.; data curation, M.S.; writing—original draft preparation, M.S.; writing—review and editing, K.Z.; visualization, M.S.; supervision, K.Z.; project administration, K.Z.; funding acquisition, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Science and Technology Projects in Yunnan Province (202202AD080013), Innovative Research Group Project of the National Natural Science Foundation of China (No.61971208), Ten Thousand Talent Plans for Young Top-notch Talents of Yunnan Province (N0.201873), Major Science and Technology Projects in Yunnan Province (202002AB080001-8), and Photonics Fund Class B (20220202, ghfund202202022131). This work is supported by Yunnan Key Laboratory of Computer Technologies Application.

Data Availability Statement: Not applicable.

Acknowledgments: We thank our lab teachers and students for their support in this research.

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

References

1. Kassens-Noor, E.; Wilson, M.; Cai, M.; Durst, N.; Decaminada, T. Autonomous vs. self-driving vehicles: The power of language to shape public perceptions. *J. Urban Technol.* **2021**, *28*, 5–24. [CrossRef]
2. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* **2020**, *37*, 362–386. [CrossRef]
3. Zuo, L.; He, P.; Zhang, C.; Zhang, Z. A robust approach to reading recognition of pointer meters based on improved mask-RCNN. *Neurocomputing* **2020**, *388*, 90–101. [CrossRef]
4. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 10–16 October 2016; pp. 21–37.
6. Huang, Z.; Wang, J.; Fu, X.; Yu, T.; Guo, Y.; Wang, R. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. *Inf. Sci.* **2020**, *522*, 241–258. [CrossRef]
7. Li, L.; Ota, K.; Dong, M. Deep learning for smart industry: Efficient manufacture inspection system with fog computing. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4665–4673. [CrossRef]
8. Xing, H.; Xiao, Z.; Qu, R.; Zhu, Z.; Zhao, B. An Efficient Federated Distillation Learning System for Multi-task Time Series Classification. *arXiv* **2021**, arXiv:2201.00011.
9. Wang, K.; Yang, J.; Yuan, S.; Li, M. A lightweight network with attention decoder for real-time semantic segmentation. *Vis. Comput.* **2022**, *38*, 2329–2339. [CrossRef]
10. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-former: Bridging mobilenet and transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 16–24 June 2022; pp. 5270–5279.
11. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
12. Li, X.; He, M.; Liu, Y.; Luo, H.; Ju, M. SPCS: A spatial pyramid convolutional shuffle module for YOLO to detect occluded object. *Complex Intell. Syst.* **2022**, 1–15. [CrossRef]
13. Adarsh, P.; Rathi, P.; Kumar, M. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Chengdu, China, 23–23 April 2020; pp. 687–694.
14. Wang, J.; Gao, Z.; Zhang, Y.; Zhou, J.; Wu, J.; Li, P. Real-Time Detection and Location of Potted Flowers Based on a ZED Camera and a YOLO V4-Tiny Deep Learning Algorithm. *Horticulturae* **2021**, *8*, 21. [CrossRef]
15. Ning, C.; Menglu, L.; Hao, Y.; Xueping, S.; Yunhong, L. Survey of pedestrian detection with occlusion. *Complex Intell. Syst.* **2021**, *7*, 577–587. [CrossRef]
16. Zerhouni, E.; Lányi, D.; Viana, M.; Gabrani, M. Wide residual networks for mitosis detection. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI), Melbourne, VIC, Australia, 18–21 April 2017; pp. 924–928.
17. Hafiz, A.M.; Parah, S.A.; Bhat, R.U.A. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. *arXiv* **2021**, arXiv:2106.07550.
18. Elayaperumal, D.; Joo, Y.H. Robust visual object tracking using context-based spatial variation via multi-feature fusion. *Inf. Sci.* **2021**, *577*, 467–482. [CrossRef]
19. Tian, J.; Yuan, W.; Tu, Y. Image compressed sensing using multi-scale residual generative adversarial network. *Vis. Comput.* **2022**, *38*, 4193–4202. [CrossRef]
20. Soviany, P.; Ionescu, R.T. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In Proceeding of the 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Linz, Austria, 20–23 September 2018; IEEE: New York, NY, USA, 2018; pp. 209–214.
21. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-stage multi-person pose machines. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 6951–6960.
22. Hafiz, A.M.; Bhat, G.M. A survey on instance segmentation: State of the art. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 171–189. [CrossRef]

23. Yi, Z.; Yongliang, S.; Jun, Z. An improved tiny-yolov3 pedestrian detection algorithm. *Optik* **2019**, *183*, 17–23. [CrossRef]
24. Oh, D.; Shin, B. Improving evidential deep learning via multi-task learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Pomona, CA, USA, 24–28 October 2022; pp. 7895–7903.
25. Sengupta, A.; Ye, Y.; Wang, R.; Liu, C.; Roy, K. Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **2019**, *13*, 95. [CrossRef]
26. Murthy, C.B.; Hashmi, M.F.; Keskar, A.G. Optimized MobileNet+ SSD: A real-time pedestrian detection on a low-end edge device. *Int. J. Multimed. Inf. Retr.* **2021**, *10*, 171–184. [CrossRef]
27. Rogelio, J.; Dadios, E.; Bandala, A.; Vicerra, R.R.; Sybingco, E. Alignment control using visual servoing and mobilenet single-shot multi-box detection (SSD): A review. *Int. J. Adv. Intell. Inform.* **2022**, *8*, 97–114. [CrossRef]
28. Shao, Z.; Cheng, G.; Ma, J.; Wang, Z.; Wang, J.; Li, D. Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. *IEEE Trans. Multimed.* **2021**, *24*, 2069–2083. [CrossRef]
29. Zhang, C.; Rameau, F.; Lee, S.; Kim, J.; Benz, P.; Argaw, D.M.; Kweon, I.S. Revisiting Residual Networks with Nonlinear Shortcuts. In Proceedings of the 30th British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019; p. 12.
30. Zhang, C.; Benz, P.; Argaw, D.M.; Lee, S.; Kim, J.; Rameau, F.; Kweon, I.S. Resnet or densenet? Introducing dense shortcuts to resnet. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Online, 1–5 January 2021; pp. 3550–3559.
31. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 311–368. [CrossRef]
32. Wang, M.; Chen, H.; Li, Y.; You, Y.; Zhu, J. Multi-scale pedestrian detection based on self-attention and adaptively spatial feature fusion. *IET Intell. Transp. Syst.* **2021**, *15*, 837–849. [CrossRef]
33. Yu, B.; Yang, L.; Chen, F. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3252–3261. [CrossRef]
34. Zeng, K.; Ma, Q.; Wu, J.; Xiang, S.; Shen, T.; Zhang, L. NLFFTNet: A non-local feature fusion transformer network for multi-scale object detection. *Neurocomputing* **2022**, *493*, 15–27. [CrossRef]
35. Lin, X.; Zhao, C.; Zhang, C.; Qian, F. Self-attention-guided scale-refined detector for pedestrian detection. *Complex Intell. Syst.* **2022**, *8*, 4797–4809. [CrossRef]
36. Zhu, X.; Guo, K.; Ren, S.; Hu, B.; Hu, M.; Fang, H. Lightweight image super-resolution with expectation-maximization attention mechanism. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1273–1284. [CrossRef]
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
38. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef]
39. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2636–2645.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
41. Cai, J.; Hu, J. 3D RANs: 3D residual attention networks for action recognition. *Vis. Comput.* **2020**, *36*, 1261–1270. [CrossRef]
42. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
43. Gosaye, K.; Moloo, R.K. A Mobile Application for Fruit Fly Identification Using Deep Transfer Learning: A Case Study for Mauritius. In Proceedings of the 2022 International Conference for Advancement in Technology, Goa, India, 21–23 January 2022; pp. 1–5.
44. Kang, S.; Park, H. Hierarchical CNN-Based Senary Classification of Steganographic Algorithms. *J. Korea Multimed. Soc.* **2021**, *24*, 550–557.