Mathematical Biosciences and Engineering

MBE, 19(12): 12833–12851.
DOI: 10.3934/mbe.2022599
Received: 05 May 2022
Revised: 05 August 2022
Accepted: 15 August 2022
Published: 01 September 2022

http://www.aimspress.com/journal/MBE

Research article

Weakly perceived object detection based on an improved CenterNet

Jing Zhou[1,*], Ze Chen[1] and Xinhan Huang[2]

[1] School of Artificial Intelligence, Jianghan University, Wuhan 430056, China
[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

* **Correspondence:** Email: zhj131@jhun.edu.cn.

Abstract: Nowadays, object detection methods based on deep neural networks have been widely applied in autonomous driving and intelligent robot systems. However, weakly perceived objects with a small size in the complex scenes own too few features to be detected, resulting in the decrease of the detection accuracy. To improve the performance of the detection model in complex scenes, the detector of an improved CenterNet was developed via this work to enhance the feature representation of weakly perceived objects. Specifically, we replace the ResNet50 with ResNext50 as the backbone network to enhance the ability of feature extraction of the model. Then, we append the lateral connection structure and the dilated convolution to improve the feature enhancement layer of the CenterNet, leading to enriched features and enlarged receptive fields for the weakly sensed objects. Finally, we apply the attention mechanism in the detection head of the network to enhance the key information of the weakly perceived objects. To demonstrate the effectiveness, we evaluate the proposed model on the KITTI dataset and COCO dataset. Compared with the original model, the average precision of multiple categories of the improved CenterNet for the vehicles and pedestrians in the KITTI dataset increased by 5.37%, whereas the average precision of weakly perceived pedestrians increased by 9.30%. Moreover, the average precision of small objects (AP_S) of the weakly perceived small objects in the COCO dataset increase 7.4%. Experiments show that the improved CenterNet can significantly improve the average detection precision for weakly perceived objects.

Keywords: object detection; anchor-free; CenterNet; attention mechanism; multi-scale feature enhancement

1. Introduction

With the popularization of artificial intelligence technology, automatic driving technology continues to develop rapidly [1]. For automatic driving systems, object detection technology plays a vital role in environmental awareness tasks [2]. Nowadays, the traditional object detection algorithms based on handcrafted features are being gradually replaced by detection technology based on a deep neural network.

Region-Convolutional Neural Network (R-CNN) is the first object detection framework based on the application of convolutional neural networks (CNN) [3], and it improves the ability of feature representation via CNN operation for detection. The faster R-CNN [4] generates regional proposals via a region proposal network (RPN) and introduces an anchor mechanism to regress the objects; it establishes the framework of the anchor-based detection algorithm to improve the detection performance. Since the anchor mechanism has been proposed, it has gradually played a critical role in popular detectors, such as YOLOv2 [5], YOLOv3 [6], YOLOv4 [7], Libra R-CNN [8] and Cascade RCNN [9].

However, owing to the development of object detection technology, the drawbacks of the anchor mechanism cannot be ignored. For instance, the detection performance of the anchor-based model is greatly affected by the size, aspect ratio and number of anchor boxes [10]. Moreover, due to the fixed size and aspect ratio of the anchor boxes, it is difficult for the anchor-based models to detect objects with large-scale variations; thus, the models need to reset the anchor boxes with different sizes and aspect ratios for the new detection task. In addition, the anchor-based models need to put dense anchor boxes on the input image to obtain a higher recall rate, which brings large amounts of hyperparameters and increases the computational complexity of the model [11]; meanwhile, most anchor boxes are considered as negative samples and only a few are considered as positive samples, which aggravates the imbalance between positive and negative samples.

Therefore, object detection algorithms that are anchor-free have attracted lots of attention in recent years, and they do not rely on pre-set anchor boxes. Compared with the traditional anchor-based method, the anchor-free detector with the simpler structure has no hyperparameters related to the anchor boxes, and it has the potential to surpass anchor-based methods in detection speed and accuracy.

The anchor-free detectors are usually divided into key point-based methods and center-based approaches. The key point-based approach first locates pre-defined or self-learned key points and then generates bounding boxes to detect objects, such as CPNDet [12], RepPoints [13] and YOLOX [14]. And, the center-based method regards the central area (center point or area) of the object as the foreground area and then predicts their distance to the four sides of the object, such as LSNet [15], GA-RPN [16] and FSAF [17].

Though those approaches achieve great detection performance, the detection performance falls when confronting complex scenes with lots of weakly perceived small objects. Therefore, we propose a novel and effective detector based on the anchor-free mechanism to detect weakly perceived objects accurately. Concretely, we first replace the backbone ResNet50 of CenterNet with ResNext50 to acquire various levels of features, and then we integrate the multi-level features of the bottom-up pathway into the corresponding features with the same scale in the top-down pathway by using the lateral connections of a feature pyramid network (FPN) [18], obtaining plentiful information on weakly sensed small objects. Simultaneously, we add the dilated encoder following the input features in the top-down pathway to enlarge the receptive fields of the weakly sensed small objects. Finally, we append the squeeze-and-excitation (SE) attention module in the detection head to enhance the key point knowledge of the weakly sensed objects.

In summary, the main contributions in this work can be summarized as follows:

1) We propose an improved anchor-free detector based on the CenterNet to elevate the detection performance for weakly perceived objects in complicated scenes.

2) We improve the feature enhancement layer of the CenterNet by adding the lateral connection structure and the dilated convolution to enrich the information of the features and amplify the reception fields for weakly sensed small objects.

3) We modify the detection head of the CenterNet with the SE attention module to enhance the key information of weakly sensed objects with small sizes.

2. Related Work

Anchor-free detectors. The anchor-free detectors require no pre-set anchor box, which makes the network structure simpler and the model more generalized. Anchor-free detectors mainly consist of two streams of center-based detectors [19] and key point-based detectors. The center-based detectors use the central region or the central point of the object to determine the positive sample and then regress its distance to the bounding box. The typical center-based works are FCOS [20], SAPD [21], etc. Another stream of key point-based detectors identifies the key points of the object to regress the bounding boxes of the objects. For instance, CornerNet [22] converts the object detection problem into the detection of a pair of key points for the object without anchors and then uses a top-left corner and a bottom-right corner of the object to predict the bounding box of the object. Referring to CornerNet, ExtremeNet [23] detects four extreme points (top-most, left-most, bottom-most, right-most) and one center point of the object to generate the bounding box. Different from the above models that detect multiple key points, CenterNet [24] regards the object as a key point to determine the center coordinates of the object through Gaussian operation and then regresses the size and position of the object.

Multi-scale feature-enhancement methods. SSD [25] is the first detector to adopt the multi-scale feature-enhancement method and multi-level feature stratification to detect objects of various sizes, and it allocates multi-scale objects to corresponding feature layers according to the size of the object. Each layer is responsible for the prediction of the object with the corresponding scale. The features of shallow layers with more detailed information are suitable for learning small objects. And, the features of deep layers with more global semantic information are appropriate for predicting large objects. DSSD [26] supposes that the insufficient semantic information and plenty of noise derived from shallow layers in SSD will weaken the classification ability of the detection network. Thus, DSSD adopts ResNet101 to integrate the global semantic information of deep layers into shallow layers. Liu et al. [27] presented the RFBNet based on SSD with a receptive field module to improve the detection performance of weakly sensed small objects. The receptive field module is composed of multi-branch convolution and expansion convolution, which expands the width of the network and enhances the adaptability of the network to multi-scale objects.

Attention mechanism. Since each channel of the feature contributes differently to the detection performance, Hu et al. [28] proposed a channel-attention model, SENet, to learn the weights of different channels, leading to the network paying more attention to the key channel information by weighting. Inspired by the SENet, ECANet [29] adopts 1D convolution to implement the local cross-channel interaction and maintain the detection performance while reducing the parameters of the model. Different from SENet and ECANet, CBAM [30] exploits both spatial- and channel-wise attention mechanisms to heighten the focus of important parts, and it contains two main components: a channel

attention module and spatial attention module. The channel attention module pays attention to the important channels of the feature, and the spatial attention module focuses on the key location information of objects. Mnih et al. [31] imported the attention mechanism to extract more small-scale features to strengthen the focus of small-scale objects and improve the detection performance of small objects.

3. CenterNet

As an anchor-free detector, CenterNet directly predicts the category and coordinates of the object on feature maps without numerous pre-setting anchor boxes, leading to fewer hyperparameters. In addition, the CenterNet determines center points via key point estimation and then regresses the object properties of location and size.

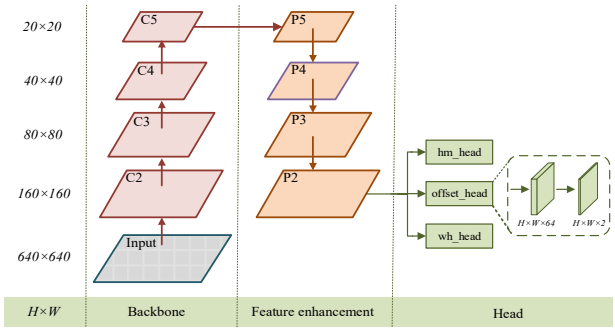


Figure 1. Architecture of CenterNet.

As shown in Figure 1, the CenterNet is composed of three parts: backbone network, feature enhancement layer and detection head. First, CenterNet extracts the preliminary features from the input image via the backbone network, and then it uses a feature enhancement layer to strengthen the semantic contexts of the features to obtain high-resolution features. Finally, the high-resolution features are applied to classify and regress in the detection head to predict the bounding boxes of objects.

4. Improvements of CenterNet

4.1. Backbone network

The standard CenterNet model takes ResNet50 as the backbone network for feature extraction. ResNet introduces residual learning into the deep network, which brings the shallow information into the deep layers of the network by performing an identity mapping operation, solving the problem of deep network degradation. However, the feature extraction effect of ResNet50 is still insufficient; thus, the grouped convolution [32] is introduced into ResNext50 [33] to extract multiple levels of features.

Concretely, ResNext50 divides the input feature into several groups and applies the block constituted with several 1 × 1 and 3 × 3 convolutions to update each group feature; then, these updated features are concatenated to enhance the feature information and the shortcut connection referring to the ResNet structure is performed to prevent network degradation. The whole construction schemes of the ResNet50 and ResNext50 are illustrated in Figure 2, where Figure 2a) is the block of ResNet50

and Figure 2b) is the block of ResNext50. As shown in Figure 2b), ResNext50 divides input features into 32 groups by grouped convolution to widen the network and applies the 1 × 1 convolution operation to the grouped features to reduce its channel dimensions; then, it adds one 3 × 3 convolution to refine semantic contexts and a 1 × 1 convolution to raise the channel dimensions. Finally, the features of each group and the original input features are aggregated to get the output feature of the residual block.

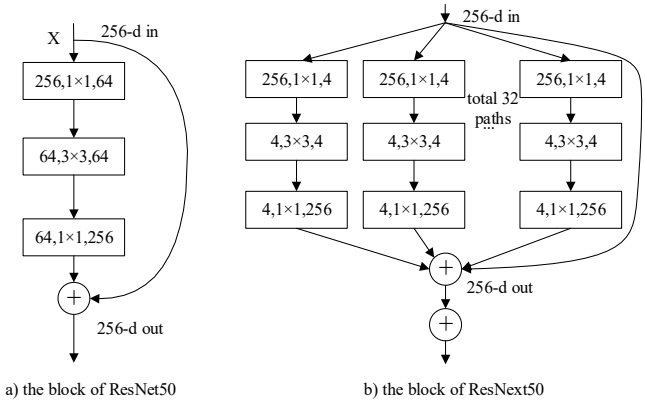


Figure 2. Comparison blocks of ResNet50 and ResNext50.

ResNext50 extracts the rich features of different levels of the network by increasing the network width, which improves the performance of the network, keeping the parameters at the same level as ResNet50. Meanwhile, considering that weakly sensed objects require sufficient information for detection, we employ the ResNext50 as the backbone network to improve the feature-extraction capability of the CenterNet model.

4.2. Improvement of the feature-enhancement layer

After obtaining the preliminary feature $C_5 \in R^{W_1 \times H_1 \times C_1}$ from the ResNext50 backbone, the feature-enhancement layer of the basic CenterNet in Figure 1 enhances the preliminary feature in three upsampling layers to generate a high-resolution feature $P_2 \in R^{W_2 \times H_2 \times C_2}$. However, the high-resolution feature P_2 is only originated from the preliminary feature C_5, and the stacked sampling layers of the basic CenterNet in Figure 1 will cause information loss, resulting in insufficient detailed information on weakly sensed objects in the feature map P_2. To solve the problem, inspired by the FPN structure, we improved the structure of the feature-enhancement layer with lateral connections to aggregate the features with different scales in the bottom-up pathway and achieve abundant detailed knowledge of weakly perceived objects in the top-down pathway. Then, we integrate the dilated encoder module [34] to enlarge the receptive field of the features and acquire more global semantic contexts for weakly sensed objects.

The architecture of the improved feature-enhancement layer is shown in Figure 3; it involves a bottom-up pathway, a dilated encoder, a top-down pathway and the lateral connections referring to the FPN structure.

The bottom-up pathway with numerous convolution layers is the feed-forward computation of

the ResNext50 backbone, which computes a feature hierarchy consisting of the features with different scales. We deem that the layers producing output maps of the same size belong to the same network stage. Thus, for the bottom-up pathway on the ResNext50 backbone, we define the four stages of S2, S3, S4 and S5 according to the sizes of the output maps. And, we denote the output features of each stage in {S2, S3, S4, S5} as C_2, C_3, C_4, C_5, and they are input to the top-down pathway of the improved CenterNet model, forming a multiple-in structure, which is different from the top-down pathway of the original CenterNet model with the single-input structure. The multiple-in structure of the improved CenterNet with the multi-scale features of C_2, C_3, C_4, C_5 can merge various input features for the top-down pathway to obtain sufficient knowledge.

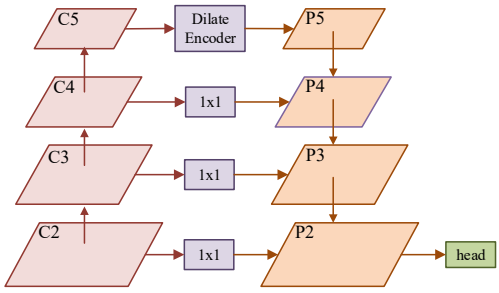


Figure 3. Structure of the improved feature-enhancement layer.

Similar to the bottom-up pathway, we detect four feature maps with different scales in the top-down pathway and define them as $\{P_2, P_3, P_4, P_5\}$. P_5 is first produced by the preliminary feature map C_5 of the ResNext50 backbone, followed by a dilated encoder, which enlarges the receptive field of Feature C_5 to acquire more context from the semantic information. And then, following the FPN structure, we merge the feature maps $\{C_2, C_3, C_4\}$ of the bottom-up pathway into the features with the same spatial size in the top-down pathway to enrich the detailed information of the output feature in the top-down pathway. Specifically, the bottom-up features C_2, C_3 and C_4, followed by a 1×1 convolution, are integrated into the corresponding feature maps in the top-down pathway via the top-down connections of the FPN, producing features of P_2, P_3 and P_4, as shown in Figure 3. With the adoption of the dilated encoder and lateral connections, the detailed information of the features with various scales in the top-down pathway is strengthened and the corresponding receptive fields are enlarged. As a result, the feature P_2 with a large receptive field and rich contextual semantic global knowledge is obtained as the output feature of the top-down pathway.

The dilated encoder in the top-down pathway contains two main components: the projector and the residual blocks, as shown in Figure 4. The projection layer first applies a 1×1 convolution to reduce the channel dimensions, and utilizes a 3×3 convolution to refine semantic contexts. Then, the output of the projector is fed into the residual blocks consisting of 1×1 convolutions and 3×3 dilated convolutions with different dilation rates to obtain the multiple receptive fields. Specifically, in each residual block, the channel dimension of the projector output is reduced via a 1×1 convolution, and a 3×3 dilated convolution is applied to improve the receptive field of the features; then, a 1×1 convolution is adopted to increase the channel dimensions. And then, the four successive dilated residual blocks are stacked to generate output features with multiple receptive fields.

Meanwhile, to reduce the calculation cost of the network, we retain the output structure of the feature-enhancement layer of the original CenterNet model and output only Feature P_2 for the

classification and regression of the detection head; then, the non-maximal suppression and other post-processing steps are eliminated to reduce the calculation cost of the model.

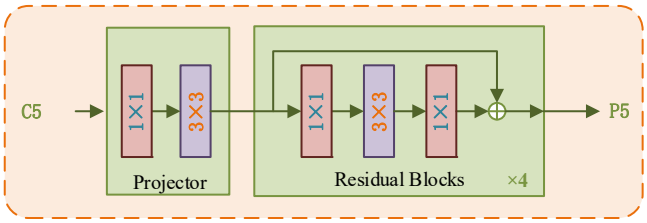


Figure 4. Dilated encoder architecture.

4.3. Embedding of attention mechanism

In the task of object detection for weakly perceived objects, the texture features play an important role. To strengthen the network's attention to the texture features of weakly sensed objects, we applied the channel attention mechanism [28] of SE module shown in Figure 5 to the CenterNet model.

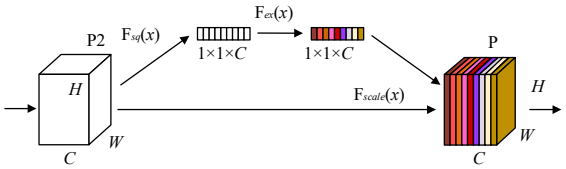


Figure 5. Attention mechanism.

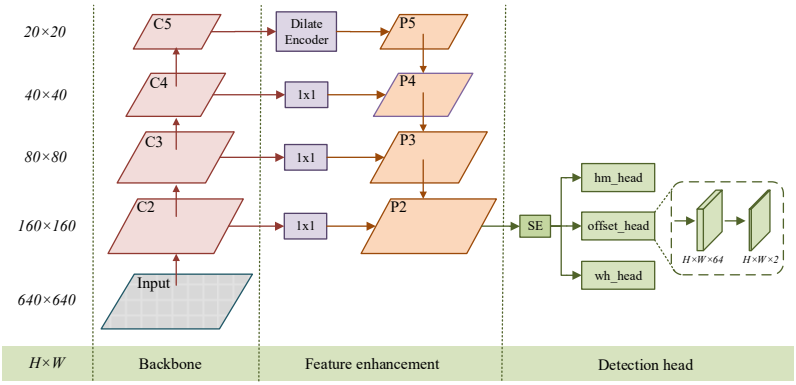


Figure 6. Architecture of the improved CenterNet.

Concretely, the output feature $P_2 \in \mathrm{R}^{W_2 \times H_2 \times C_2}$ with a width W_2 and height H_2 from the feature-enhancement layer is input to the SE module. And, the feature P_2 is first updated by a global pooling operation to produce an embedding vector representing the global distribution of channel feature responses. Then, the number of channels of the embedding vector is decreased via a 1 × 1 convolution to reduce the computations. The full connection layer is utilized to keep the number of channels of the embedding vector consistent with the number of channels of the input feature P_2. Subsequently, the

sigmoid function takes the embedding vector as input and produces a series of channel-wise weights. We multiply these weights with the input feature P_2 to generate the attention-weighted feature P as the output of the SE block to guide the network to focus on key channels; the overall structure of the improved CenterNet is shown in Figure 6.

Finally, the attention-weighted feature P is input to the detection head for classification and regression to generate the refined object bounding boxes. The detection head includes three branches: a heatmap branch, an offset branch and the size branch. For the heatmap branch, the key point heatmaps are generated by applying the Gaussian operation to the feature P, and then the heatmaps are used to predict the categories of objects and center coordinates of objects. For the offset branch, the offsets of the center coordinates are estimated. And, the sizes of the bounding boxes of objects are regressed in the size branch.

4.4. Loss function

The improved CenterNet proposed in this work is trained with the total loss, which consists of three parts: the key point loss (L_k), offset loss (L_{off}) and size loss (L_{size}), corresponding to the three branches of the detection head, respectively. And, the total loss is formulated as

$$L_{det} = L_k + 0.1L_{size} + L_{off} \tag{1}$$

The key point loss is realized via the key point heatmaps to learn the categories and center coordinates for objects, as folllows:

$$L_k = \frac{-1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right) & if \quad Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha}\log\left(1-\hat{Y}_{xyc}\right) & otherwise\end{cases} \tag{2}$$

where Y_{xyc} represents the ground-truth feature value at the (x, y) coordinates for Category C on the key point heatmaps, which is generated by Gaussian operation on the input feature of the heatmap branch, and $\hat{Y}_{xyc}$ represents the predicted value obtained via the network, while the prediction $\hat{Y}_{xyc}=1$ represents a detected key point taken as the center point at the (x,y) coordinates for Category C; $\hat{Y}_{xyc}=0$ denotes the value of the background point. Meanwhile, α and β are the hyperparameters of the focal loss, where α is 2 and β is 4. N is the number of key points in the heatmaps.

Further, to recover the bias caused by the downsampling of the convolution operation, we utilize the offset loss to learn the offsets of the center coordinates, and it is constructed with L_1 loss, as follows:

$$L_{off} = \frac{1}{N}\sum_{p}\left|\hat{O}_{p^*} - \left(\frac{p}{R} - p^*\right)\right| \tag{3}$$

where P is the center of the original input image of the network, p^* is the center of the scaled-down feature map achieved by the convolution operation and R is the stride of the convolution operation; then, we can obtain the real bias of $\left(\frac{p}{R}-p^*\right)$, and the bias of network prediction is illustrated as $\hat{O}_{p^*}$.

In addition, for the k-th object, its bounding box is denoted as $(x_1^{(k)}, y_1^{(k)}, x_2^{(k)}, y_2^{(k)})$, and we can get the ground-truth object size of $S_k = (x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)})$; then, the size loss based on the L_1 loss function is applied to regress the size of the object bounding boxes, as follows:

$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{p_k} - s_k\right| \tag{4}$$

where $\hat{S}_{pk}$ is the predicted regression value.

5. Experiment

5.1. Dataset

We evaluate our proposed weakly sensed object detection network on the KITTI dataset [35] and COCO dataset [36], respectively. KITTI is one of the popular datasets in the autonomous driving field and it provides a large number of images of complex environments, including urban areas, roads, rural scenes, etc. The KITTI dataset contains a total of 7481 images, including 33,252 vehicles and 6340 pedestrians. There are eight categories in the KITTI dataset, including Car, Van, Truck, Pedestrian, Person-sitting, Cyclist, Tram and Misc. The fact is that pedestrians have a small size and lack of abundant feature information; thus, they are usually regarded as weakly sensed small objects. At the same time, lots of weakly sensed small vehicles caused by distance, truncation and occlusion are difficult to be perceived and detected. Therefore, we chose two categories in the KITTI dataset, including pedestrians (Pedestrian, Person-sitting and Cyclist) and vehicles (Car, Van, Truck and Tram) to evaluate the effectiveness of our improved CenterNet for weakly perceived objects. Meanwhile, we divided the dataset into the training set, the verification set and the test set in the proportions of 0.8, 0.1 and 0.1, respectively.

Different from the KITTI dataset, the MS COCO dataset has more samples, providing 330,000 images, 1.5 million object instances and 80 object categories. The images in the COCO dataset were mainly captured from complex daily scenes in real environments, containing numerous objects of various types; each image contains 3.5 object categories on average. Meanwhile, COCO divides objects into three scales: the large objects with sizes greater than 96 × 96 pixels, the medium objects with sizes ranging from 32 × 32 pixels to 96 × 96 pixels and the small objects with sizes less than 32 × 32 pixels to measure the average precision (AP) of multiple categories (mAP) values, respectively, where the large objects account for 24%, the medium objects account for 34% and the small objects account for 41%; that is, almost half of the objects in the COCO dataset are small, and they lack sufficient information to be perceived. Therefore, it is recommended to leverage the MS COCO dataset containing numerous small objects to evaluate our detection performance of weakly sensed objects. To clearly demonstrate the improvement of the detection performance of our proposed detector for weakly perceived objects, we evaluated our detector on the large objects (L), the medium objects (M) and the small objects (S) in the COCO dataset, respectively; we discuss the ability of our detector to detect weakly sensed objects based on the detection accuracy of the small objects (S).

5.2. Training details and evaluation metrics

We trained our proposed model in an end-to-end manner with the Adam optimizer; the pre-training weights of ResNet50 and ResNext50 were obtained from ImageNet.

The overall training process includes the frozen stage and unfrozen stage. During the first 50 epochs of training, the parameters of the backbone in the network are frozen and will not be updated,

while the parameters of the other parts of the network are updated. For the KITTI dataset, we trained the entire network with a batch size of 8 and learning rate of 0.001 using a GTX 1660Ti GPU. For the COCO dataset, the network was trained with a batch size of 16 and learning rate of 0.001 using two GTX 2080 Ti GPUs. After 50 epochs, the parameters of the overall network were updated. For the KITTI dataset, the network was updated with a batch size of 4 and learning rate of 0.0001. For the COCO dataset, the network was trained with a batch size of 8 and learning rate of 0.0001.

To illustrate the effectiveness of our proposed detection network, we adopted the evaluation metrics of the AP and mAP to evaluate the performance of the model. The AP is the average precision of a single category, and it is the index to measure the model performance. And, mAP is the average value of the AP for multiple categories; it measures the performance of the model for all categories. The AP is denoted as

$$\mathrm{AP} = \int_0^1 \mathrm{u}(v)\, d\mathrm{v} \tag{5}$$

where u demonstrates accuracy and v denotes the recall rate.

5.3. Evaluation and analysis

To select the appropriate components to compose our proposed improved CenterNet, we conducted a series of experiments using the KITTI dataset; then, we trained our proposed network and evaluated it on the KITTI and COCO datasets, respectively.

Backbone. For the selection of a suitable backbone network, we compared the performance of CenterNet on ResNext50 and ResNet50, respectively. As shown in Table 1, compared with ResNet50, the mAP of the CenterNet on ResNext50 was increased by 1.81%, achieving gains of 0.96% for vehicles and 2.66% for the weakly perceived small objects of pedestrians. Obviously, the performance of the CenterNet on ResNext50 was better than that of the CenterNet on ResNet50. Thus, we selected the ResNext50 as the backbone for further experiments.

Table 1. Comparison of the models on ResNet50 and ResNext50 using the KITTI dataset. The evaluation metric is the APm with an Intersection over Union (IoU) threshold of 0.5 for pedestrians and vehicles.

Method	Backbone	AP (pedestrian)	AP (vehicle)	mAP
CenterNet	ResNet50 [37]	72.70	93.31	83.01
CenterNet	ResNext50 [33]	75.36	94.27	84.82

Table 2. Comparison of the models with different feature-enhancement layers ("+": with improved feature-enhancement layer, "-": original feature-enhancement layer).

Backbone	Feature enhancement layer	AP (pedestrian)	AP (vehicle)	mAP
ResNext50	-	75.36	94.27	84.82
ResNext50	+	79.30	94.86	87.08

Feature-enhancement layer. Based on the backbone of ResNext50, we compared the improved feature-enhancement layer with the original structure of the CenterNet. The experimental results shown in Table 2 indicate that the improved structure achieved 2.26% mAP gains, whereas the AP of vehicles increased by 0.59% and the AP of the weakly perceived small objects of pedestrians increased

by 3.94%. It can be seen that the detection accuracy of the weakly perceived small objects of pedestrians has been significantly promoted by improving the feature enhancement layer.

Attention module. According to the results in Table 2, we selected ResNext50 with the improved feature-enhancement layer as a baseline to append various attention modules for comparative experiments. Specifically, we added the SE module, convolutional block attention module (CBAM) and efficient channel attention (ECA) module into the detection head of the baseline, respectively, to verify the effect of the attention mechanism on the model. As displayed in Table 3, the AP of the weakly perceived small objects of pedestrians of the CenterNet with the SE module ranked first among all of the models. Moreover, compared with the baseline model, the mAP of the CenterNet model with the SE module was increased by 1.30%, where the AP of the weakly perceived pedestrians achieved a 2.7% increase, but the AP of vehicles slightly dropped.

Table 3. Comparison of models with improved feature-enhancement structure and attention modules ("×": without the attention module).

Backbone	Feature-enhancement layer	Attention	AP (pedestrian)	AP (vehicle)	mAP
ResNext50	+	×	79.30	94.86	87.08
ResNext50	+	SE [28]	82.00	94.77	88.38
ResNext50	+	CBAM [30]	80.65	95.09	87.87
ResNext50	+	ECA [29]	78.39	94.31	86.35

According to the above discussion, we propose our improved CenterNet consisting of the backbone of ResNext50, the improved feature-enhancement layer and the detection head with the SE module. Meanwhile, the pedestrians are usually small in size and possess too little information to be perceived, resulting in hard detection; a series of detectors with multi-scale structures are proposed to strengthen the features of the small weakly sensed objects for detection, such as FCOS, YOLOV4, YOLOV3, YOLOX and SSD. Therefore, to verify the effect of our proposed detector for the weakly sensed objects of pedestrians from the KITTI dataset, we compared our proposed improved CenterNet with different state-of-the-art detectors of multi-scale structure, such as the anchor-free detectors of FCOS and YOLOX, the representative anchor-based detectors of SSD, YOLOV3 and YOLOV4 and the latest anchor-based algorithms of YOLOV3-SPP, YOLOV3-SPP-ASFF, YOLOV3-SPP-ASFF-SE and ResNext-SSD; the results are shown in Table 4.

Table 4. Performance comparison for various algorithms on the KITTI dataset. The evaluation metric is the AP, with an IoU threshold of 0.5 for pedestrians and vehicles.

Detector	Backbone	AP (pedestrian)	AP (vehicle)	mAP
FCOS [20]	ResNet50	71.63	93.63	82.63
YOLOV3 [6]	Darknet53	72.80	93.90	83.35
YOLOV3-SPP [38]	Darknet53	78.10	95.80	86.95
YOLOV3-SPP-ASFF [38]	Darknet53	79.70	95.50	87.60
YOLOV3-SPP-ASFF-SE [38]	Darknet53	80.00	95.60	87.80
YOLOV4 [7]	CSPDarknet53	82.40	89.93	86.16
YOLOX [14]	CSPDarknet	72.19	91.40	81.80
SSD [25]	ResNet50	69.8	86.5	78.15
ResNext-SSD [39]	ResNext50	81.10	92.40	86.75
CenterNet [24]	ResNet50	72.70	93.31	83.01
Ours	ResNext50	82.00	94.77	88.38

Table 4 shows that our improved CenterNet achieves an improvement of 5.37% for the mAP value relative to the basic CenterNet, the AP gain of vehicles was 1.46%, and the AP value for weakly sensed pedestrians was remarkably increased by 9.3%. Besides, compared with previous state-of-the-art multi-scale detectors, our proposed model obtained the highest mAP value for vehicles and pedestrians, where the AP value of weakly perceived pedestrians was higher than that for most detectors, except for YOLOV4; and, it achieved a 0.9% gain at least. Although our AP for pedestrians dropped slightly by 0.4%, contrasting with YOLOV4, the overall detection performance of the mAP value showed a 2.22% improvement. In summary, our proposed model appears to achieve excellent performance for the detection of weakly perceived small pedestrians in the KITTI dataset, which has numerous small pedestrians and certain weakly sensed vehicles of small sizes.

To further validate the effectiveness of our detector in detecting weakly perceived small objects, we evaluated our method on the objects of the three scales of large (L), medium (M) and small (S) in the COCO dataset, respectively; the experimental results are shown in Table 5. Meanwhile, Table 5 shows the detection accuracy of other state-of-the-art detectors on the three scales of objects.

Table 5. Performance comparison for various detectors on the COCO dataset. The evaluation metric is the AP, which is the average of 10 detection precision values given 10 IoU thresholds (0.5:0.05:0.95). And, the AP_S, AP_M and AP_L represent the AP of the small objects (S), medium objects (M) and large objects (L), respectively.

Detector	Backbone	AP	AP_S	AP_M	AP_L
SSD [25]	VGG16	28.8	10.9	31.8	43.5
YOLOv3 [6]	Darknet53	31.0	15.2	33.2	42.8
YOLOv3-SPP+ASFF [38]	Darknet53	38.1	16.1	41.6	53.6
YOLOv3-SPP [38]	Darknet53	36.0	20.6	37.4	46.1
YOLOv4 [7]	CSPDarknet53	43.5	26.7	46.7	53.3
RFBNet [27]	HarDNet68	33.9	14.7	36.6	50.5
ATSS [40]	ResNet101	43.6	26.1	47.0	53.6
RDSNet [41]	ResNet101	36.0	17.4	39.6	49.7
CenterNet [24]	ResNet50	40.7	20.3	47.1	54.2
Ours	ResNext50	43.9	27.7	48.2	55.4

As shown in Table 5, the AP value of our proposed improved CenterNet ranked first among all of the methods, leading to a 3.2% AP gain over the basic CenterNet. Meanwhile, for the detection of small objects, our proposed detector achieved a 7.4% AP_S gain over the standard CenterNet, and at least 1% AP_S improvement over other detectors, which demonstrates that our detectors can effectively promote the detection performance for weakly sensed objects of small size. Furthermore, for the detection of the objects with large scales and medium scales, the AP_M and AP_L of our framework have slight improvements over the basic CenterNet, achieving 1.1 and 1.2% gains, respectively. We deem that the added attention module in the detection head and the proposed improved feature-enhancement layer allow weakly sensed small objects with insufficient information to yield abundant knowledge and attract more attention for detection. By contrast, the large-scale and the medium-scale objects already have rich features; hence, our improvements on CenterNet yield little effect on these objects. Overall, the proposed detection framework in this paper effectively improves the detection precision and achieves great detection performance for weakly sensed small objects from the COCO dataset.

5.4. Ablation study

By summarizing the above experimental results, we designed the ablation experiments using the KITTI dataset to verify the effectiveness of various components in the improved CenterNet. In the ablation experiments, the backbone network, feature-enhancement layer and attention module in the detection head of the improved CenterNet were analyzed; the experimental results are shown in Table 6, where "+" indicates the model with the improved feature-enhancement layer and "√" represents the model using the SE attention module in the detection head.

Table 6 shows that the improved feature-enhancement layer can promote the performance of the original CenterNet model better than the improvements of the backbone network and SE module. Specifically, based on the backbone of ResNet50 and ResNext50, the mAP of the CenterNet model with an improved feature-enhancement layer was increased by 3.86 and 4.07%, respectively; however, the mAP of the CenterNet model with the SE module was only increased by 0.49 and 2.24%, respectively. Combining with the improved feature-enhancement layer and SE module in the detection head, for the backbone of ResNet50 and ResNext50, the improved model achieved 4.65 and 5.37% mAP gains, respectively. Meanwhile, the results in Table 5 show that the detection performances of different improved models with ResNext50 were better than those with ResNet50.

Table 6. Results of ablation study using the KITTI dataset ("√": adding the attention module). The evaluation metric is the AP, with an IoU threshold of 0.5 for pedestrians and vehicles; the maximum AP value or mAP value in each column is bolded.

Case	Backbone	Feature-enhancement layer	Attention	AP (pedestrian)	AP (vehicle)	mAP
Origin	ResNet50	-	×	72.70	93.31	83.01
Case 1	ResNet50	+	×	78.98	94.77	86.87
Case 2	ResNet50	-	√	73.61	93.40	83.50
Case 3	ResNet50	+	√	80.24	95.08	87.66
Case 4	ResNext50	-	×	75.36	94.27	84.82
Case 5	ResNext50	+	×	79.30	94.86	87.08
Case	ResNext50	-	√	75.72	94.77	85.25
Case 7	ResNext50	+	√	82.00	94.77	88.38

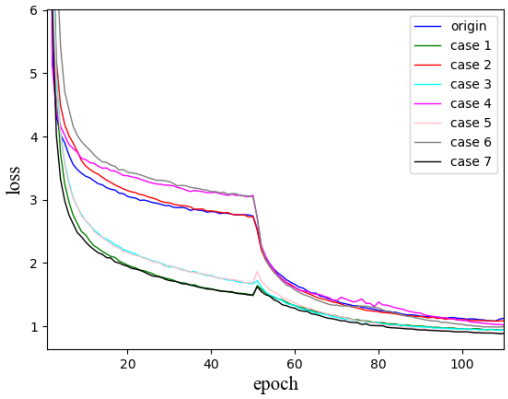
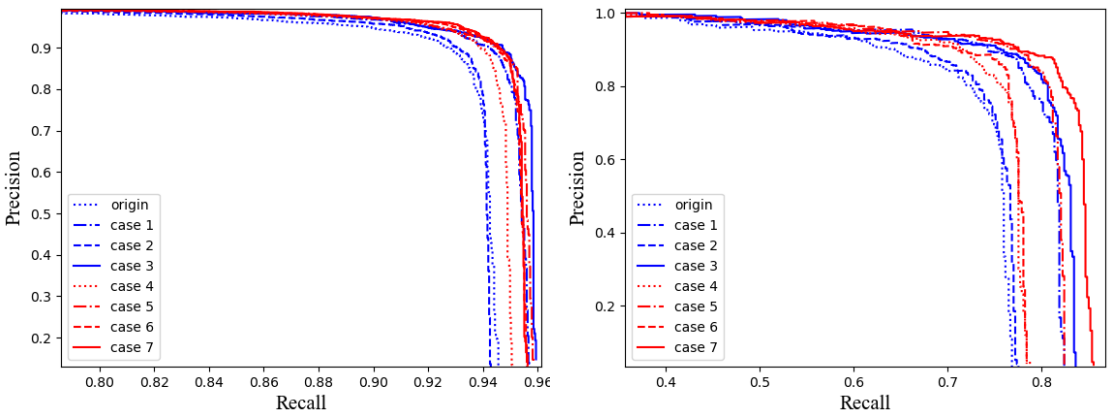


Figure 7. Comparison of training loss.

Subsequently, as shown in Figure 7, we compared the training losses for various models listed in Table 6; the loss of each improved model from Cases 1 to 7 was lower than that for the original model,

and the losses of the models with the improved feature-enhancement layer for Cases 1, 3, 5 and 7 decreased faster than other models. It proves that the improved feature-enhancement layer can effectively accelerate the rate of convergence of the CenterNet model. And. the model in Case 7 gained the least training loss, which proves its powerful learning ability.



a) PR curve of vehicle b) PR curve of pedestrian

Figure 8. Precision-recall (PR) curves for all models.

In Figure 8, we present the precision-recall (PR) curves for the various models displayed in Table 6; the PR curves for the vehicles are displayed in Figure 8 a), and the PR curves of pedestrians are shown in Figure 8b). And, all of the PR curves indicate that the precision and recall values for all models from Cases 1 to 7 listed in Table 6 have been significantly promoted relative to the original CenterNet. And, among all of the cases in Table 6, the improved model of Case 7 with the improved feature-enhancement layer and the SE block had the largest area of the PR curve and achieved the best performance.

5.5. Qualitative results

Some qualitative results for the KITTI dataset and COCO dataset that were obtained via our proposed detector and the standard CenterNet for weakly perceived objects are displayed in Figure 9. As shown in Figure 9a), our detector effectively detected the weakly perceived objects of the small pedestrians due to distance in the complex environment, while it was undetected by the original CenterNet. The weak pedestrian with poor illumination in the shadow on the left of Figure 9b) was ignored by the original model, while it could be detected by our detector. In Figure 9c), our detector could accurately detect the weakly perceived object of the occluded vehicle, which was missed by the original detector. Moreover, in the case of the COCO dataset, for the weakly sensed objects of the occluded persons of Figure 9d), the original model ignored them, while they were detected by our detector. And, in Figure 9e),f), our detector successfully located the objects with small sizes due to the occlusion or distance, which failed to be detected via the basic CenterNet.

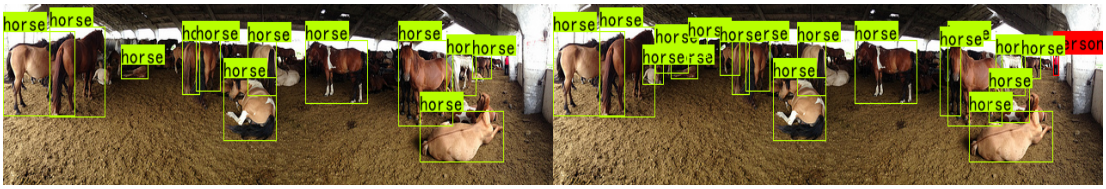
In summary, our proposed detector given as the improved CenterNet can detect weakly perceived pedestrians and other weakly perceived objects with small sizes caused by occlusion, truncation and distance.

a) Distant pedestrians with small size in the KITTI dataset

b) Small pedestrians in poor light in the KITTI dataset

c) Small vehicle with occlusion in the KITTI dataset

d) Small objects in the COCO dataset



e) Small objects with occlusion in the COCO dataset

f) Distant objects with small size in the COCO dataset

Figure 9. Qualitative results of the proposed improved CenterNet on the KITTI and COCO datasets. For the images in each group, the left plot shows the result of the original CenterNet model, and the result of our detector is displayed in the right image. The detected objects are shown with the bounding boxes.

6. Conclusions

Aiming at the problem of missed detection for weakly perceived small objects in complex

environments, we have proposed an improved CenterNet based on the anchor-free mechanism. First, ResNext50, instead of ResNet50, was adopted to be a backbone network, as it improves the ability of the feature extraction. Second, the feature-enhancement layer has been improved to strengthen the semantic information and enlarge the reception fields for the weakly sensed objects by combining the FPN structure and dilated convolution module. Finally, by appending the attention module in the detection head, the key information of the weakly sensed small objects is enhanced. The experimental results show that our improved model can elevate the detection precision of the model and accelerate the convergence speed of the original model, achieving a good effect on the detection of weakly sensed objects.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62106086) and Natural Science Foundation of Hubei Province (No. 2021CFB564).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. L. H. Wen, K. H. Jo, Deep learning-based perception systems for autonomous driving: A comprehensive survey, *Neurocomputing*, **489** (2022), 255–270. https://doi.org/10.1016/j.neucom.2021.08.155
2. X. Gao, G. Y. Zhang, Y. J. Xiong, Multi-scale multi-modal fusion for object detection in autonomous driving based on selective kernel, *Measurement*, **194** (2022), 111001. https://doi.org/10.1016/j.measurement.2022.111001
3. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014), 580–587.
4. S. Q. Ren, K. M. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in *Proceedings of the Advances in Neural Information Processing Systems*, (2015), 91–99.
5. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 7263–7271. https://doi.org/10.1109/CVPR.2017.690
6. J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, preprint, arXiv:1804.02767, https://doi.org/10.48550/arXiv.1804.02767
7. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, preprint, arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934
8. J. M. Pang, K. Chen, J. P. Shi, H. J. Feng, W. L. Ouyang, D. H. Lin, Libra r-cnn: Towards balanced learning for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 821–830.

9. Z. W. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 6154–6162.
10. T. Y. Lin, P. Goyal, R. Girshick, K. M. He, P. Dollar, Focal loss for dense object detection, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 2980–2988.
11. G. Zhao, J. Pang, H. Zhang, J. Zhou, L. J. Li, Anchor-free network for multi-class object detection in remote sensing images, in *2020 39th Chinese Control Conference (CCC)*, IEEE, (2020), 7510–7515. https://doi.org/10.23919/CCC50068.2020.9188903
12. K. Duan, L. Xie, H. Qi, S. Bai, Q. Huang, Q. Tian, Corner proposal network for anchor-free, two-stage object detection, in *Computer Vision-European Conference on Computer Vision (ECCV) 2020. Lecture Notes in Computer Science*, Springer, Cham, **12348** (2020), 399–416. https://doi.org/10.1007/978-3-030-58580-8_24
13. Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, Reppoints: Point set representation for object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9657–9666.
14. Z. Ge, S. T. Liu, F. Wang, Z. M. Li, J. Sun, Yolox: Exceeding yolo series in 2021, preprint, arXiv:2107.08430. https://doi.org/10.48550/arXiv.2107.08430
15. K. W. Duan, L. X. Xie, H. G. Qi, S. Bai, Q. M. Huang, Q. Tian, Location-sensitive visual recognition with cross-iou loss, preprint, arXiv:2104.04899. https://doi.org/10.48550/arXiv.2104.04899
16. J. Wang, K. Chen, S. Yang, C. Loy, D. Lin, Region proposal by guided anchoring, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 2965–2974.
17. C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 840–849.
18. T. Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, S. Belongle, Feature pyramid networks for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2117–2125.
19. S. Zhang, C. Chi, Y. Q. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 9759–9768.
20. Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9627–9636.
21. C. Zhu, F. Chen, Z. Shen, M. Savvides, Soft anchor-point object detection, in *Proceedings of the ECCV*, (2020), 91–107. https://doi.org/10.1007/978-3-030-58545-7_6
22. H. Law, J. Deng, CornerNet: Detecting objects as paired keypoints, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 734–750.
23. X. Zhou, J. Zhuo, P. Krahenbuhl, Bottom-up object detection by grouping extreme and center points, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 850–859.

24. X. Zhou, D. Wang, P. Krähenbühl, Objects as points, preprint, arXiv:1904.07850. https://doi.org/10.48550/arXiv.1904.07850
25. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, Ssd: Single shot multibox detector, in *Computer Vision -European Conference on Computer Vision (ECCV)*, (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
26. C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, DSSD: Deconvolutional single shot detector, preprint, arXiv:1701.06659. https://doi.org/10.48550/arXiv.1701.06659
27. S. Liu, D. Huang, Y. H. Wang, Receptive field block net for accurate and fast object detection, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 385–400.
28. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the CVPR*, (2018), 7132–7141. https://doi.org/10.48550/arXiv.1709.01507
29. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539.
30. S. H. Woo, J. C. Park, J. Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19.
31. V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Recurrent models of visual attention, *J. Adv. Neual. Inf. Process. Syst.,* **3** (2014), 2204–2212.
32. J. Shin, H. J. Kim, PresB-Net: parametric binarized neural network with learnable activations and shuffled grouped convolution, *PeerJ Comput. Sci.*, **8** (2022), e842. https://doi.org/10.7717/peerj-cs.842
33. S. Xie, R. Girshick, P. Dollár, Z. W. Tu, K. M. He, Aggregated residual transformations for deep neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1492–1500.
34. Q. Chen, Y. Wang, T. Yang, X. Zhang, You only look one-level feature, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 13039–13048.
35. A.Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. (2012), 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074
36. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Microsoft coco: Common objects in context, in *Proceeding of the European conference on computer vision (ECCV)*, (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
37. K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778.
38. Y. Li, Research of lightweight vehicle and pedestrian detection based on CNN, Master Thesis, North China University, 2021.
39. L. X. Meng, Research on vehicle pedestrian detection method based on deep learning, Master Thesis, North China University, 2021.
40. S. Zhang, C. Chi, Y. Q. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 9759–9768.

41. S. Wang, Y. Gong, J. Xing, L. Huang, C. Huang, W. Hu, RDSNet: A new deep architecture forreciprocal object detection and instance segmentation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 12208–12215. https://doi.org/10.1 609/aaai.v34i07.6902

AIMS Press