Dual-mode serial night road objectdetection model based on depth separable and self-attention mechanism

Journal:	*IEEE Transactions on Vehicular Technology*
Manuscript ID	VT-2023-01250
Suggested Category:	Regular Paper
Date Submitted by the Author:	01-Apr-2023
Complete List of Authors:	yang, qin; Xijing University, Ma, Yahong; Xijing University Zhao, Zeyu ; Xian Jiaotong University Li, Linsen; Shanghai Jiao Tong University
Keywords:	Computer Vision, Image processing, Assisted Driving
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
classification-pytorch-main.lzh	

Dual-mode serial night road objectdetection model based on depth separable and self-attention mechanism

Qin Yang, Yahong Ma*, Zeyu Zhao, Linsen Li

***Abstract*—Existing road detection at night mainly focuses on vehicle detection, but pedestrian detection is also particularly important in the complex environment of night roads. Based on the idea that one module only focuses on one task, this paper proposes a nighttime object detection method under the dual-module framework, which mainly focus on vehicle and pedestrian. In order to perform multi-scale fusion of local features and global features and realize the rapid acquisition of prediction boxes, Module 1 adopts a lightweight depthwise separable network. In Module 2, the recognition of useful features is enhanced by improving the internal structure of the residual block and increasing the self-attention mechanism, thereby improving the network performance and obtaining better classification accuracy. The efficacy of the proposed approach has been evaluated on two benchmark datasets, namely, the Berkeley Deep Drive dataset and the Hong Kong Vehicle dataset. Experimental results demonstrate that the proposed method outperforms several state-of-the-art object detection methods, exhibiting a superior accuracy rate of 93.79% under high standard IoU threshold, while maintaining a high processing speed of 42.3 frames per second (FPS). Consequently, these results verified the potential of the proposed method for advancing the field of object detection in night.**

***Index Terms*— Attention Mechanism, Lightweight, Muti-scale Fusion, Night Object Detection**

I. INTRODUCTION

In recent decades, road traffic accidents have only increased, and human errors accounted for as high as 66.8%, which is the main factor affecting road traffic accident. Vehicle errors (25.6%) and environmental factors (7.6%) are respectively secondary, and the third influencing factor [1]. And most of these accidents happen at night. Therefore, nighttime road object detection, as one of the components of advanced assistance systems and autonomous driving, is very important for road safety in intelligent transportation systems (ITS) [2].

Due to the obvious contrast between objects such as vehicles and the background on the daytime road, the features presented by vehicles or other objects are significant. Headlights are currently the feature that most nighttime vehicle detection methods focus on [3-5], because the headlights are turned on at night, and the headlights are noticed as the best feature for nighttime vehicle detection. The task of object detection in autonomous driving presents a host of challenges such as parking and moving vehicles, shadows, adverse weather conditions, lighting variations, and reflections caused by rain and snow, etc. Moreover, in many driving scenes, the headlights of other vehicles are obstructed by surrounding traffic, leading to additional difficulties. Therefore, autonomous vehicles must possess a robust object detection system that can handle diverse driving scenarios, where the detection of headlights is often imprecise. Detailed explanations of object detection techniques are available in the existing literature [6-9].

Kuang et al. [10-12] proposed a method for nighttime image augmentation using bio-inspired techniques, which was used for feature fusion and object classification. In their subsequent work, they combined traditional segmentation and deep learning features to generate regions of interest (ROI) for target detection. This approaches fall under the multi-level learning framework, but it is not an end-to-end learning framework, which can make the training process morecomplex. V et al. [15] used the classical augmentation method and the method based on CycleGAN to improve the accuracy of the night detection perception module, but the appeared visual artifacts had a great impact on the object detection. Shao et al. [16] presented a novel cascaded detection network framework, FteGanOd, that comprises two primary modules, namely, Feature Translation Enhancement (FTE) and Object Detection (OD). The FTE module is responsible for feature translation, whereas the OD module is responsible for detecting objects in the input images. However, there are too many omissions in the detection of long-distance small objects at night and high-speed road conditions, which may cause safety hazards. The current mainstream deep learning method needs a large amount of data sets for training

This work is supported by the General Projects of Shaanxi Science and Technology Plan(No.2023-JC-YB-504), the Shaanxi province innovation capacity support program (No.2018KJXX-095) and the Xijing University special talent research fund (No.XJ17T03) (*Corresponding author: Second Yhong Ma*.) The first author contributed equally.

Yang Qin studied at School of Electronic Information, Xijing University, Xi'an, China, (Email: chaorenyangqin@icloud.com).

Yahong Ma *. Works at at School of Electronic Information, Xijing University, Xi'an, China, (Email: yahongma@sina.com).

Zeyu Zhao works at the School of Computer Science, Xi'an Jiaotong University (email: 1719048836@qq.com ）

Linsen Li works at the School of Cyberspace Security at Shanghai Jiao Tong University, is a visiting scholar at the University of California, Los Angeles, and is a member of IEEE (email: lsli@sjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

and the CycleGAN is usually choose to expand the data set. This method can convert styles while maintaining the characteristics of a given input image without collecting more databases. However, there is a bias in the loss function of this method. In a dark environment, the discriminator may give high scores even if the resulting image does not meet the criteria, and the loss function may encourage the generator to generate incorrect behavior.

Through the observation of road, the main objects on the road include vehicles and pedestrians, which can be located anywhere in the image and have different sizes. Therefore, automatic generation of a window set containing various objects is an important step to improve the robustness and accuracy of road object detection. Generate a set of window detection boxes covering all vehicles and pedestrians as data to feed into the classification model. The most common object proposal method named EdgeBoxes [17] consists of a weighted sum of vehicle light detections. In Reference [18], a Bayesian saliency map-based object proposal generator was proposed, which combines multi-scale sliding windows, proposal rejection, scoring, and non-maximum suppression, but also relies too much on vehicle light feature detection and has limitations. Aiming at the above problems, a lightweight convolutional model focusing on object proposals were trained in this paper. The model proposed in this paper does not rely solely on the feature detection of vehicle light and can learn weight adjustment through regional similarity and local contrast in the nighttime road image. The framework of our proposed nighttime road object detection method is shown in Fig. 1.

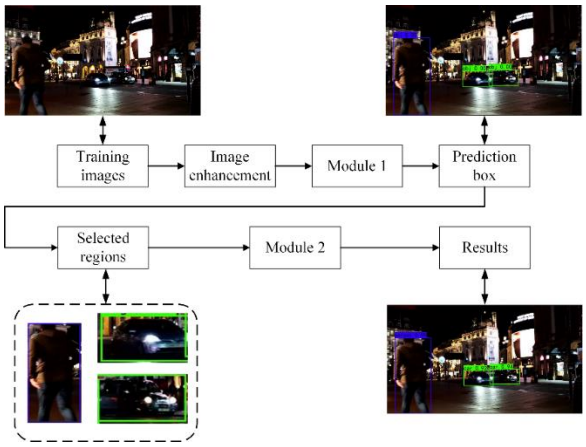

Fig. 1. Flow chart of road object detection at night

There are two modules in the flow chart of road object detection at night, shown in Fig. 1. Module 1 is a lightweight depthwise network, which is used as the object proposal box to improve the efficiency of window detection and reduce the computational burden of the entire detection process. In order to subdivide the extracted night object area Module 2 is modified on the Resnet framework and added the self-attention mechanism CBME[19]. Using the object proposal box predicted by Module 1, only the images in the object proposal box are sent to Module 2 for classification. Through Module 2, feature extraction and category classification are performed on objects in each object frame. The contributions of this paper are outlined below.

1) Aiming at the singleness of car light features in the existing night vehicle detection methods, a lightweight object proposal box model is proposed, which uses the useful features of the object in the image to select the anchor box.

2) A model specially designed for nighttime road object classification is proposed. By extracting the uniqueness of nighttime features in the object proposal box, adding a self-attention mechanism to focus on useful features, and adjusting the model structure according to training feedback to achieve excellent results performance.

3) The dual-mode series of classification after object proposal enables each module to focus on a single task object, increasing detection efficiency and accuracy, and the lightweight network and depthwise separable technology make the entire model more efficient.

The structure of this paper is as follows: Section II outlines the related work that has been conducted in the field of nighttime road object detection. Section III describes the proposed dual-mode serial object detection method in detail, which involves object proposal and classification. The object proposal method and model improvement techniques are presented in this section. The experimental results are discussed in Section IV. Lastly, Section V summarizes the conclusions drawn from this study and highlights future research directions.

II. Related work

Lights are developed as key information for vehicle object detection at night [3], and headlights and taillights containing red or bright lights are used as regions of interest. Machine learning technology has become the mainstream as nighttime object detection, some classic techniques to obtain object region proposals include threshold segmentation-based methods [20-21], saliency map-based methods [22], or using manually labeled features [23]. Spatial clustering technology with automatic multi-level threshold segmentation is also used to detect vehicles at night [5]. O et al. [24] used the color space of HSV to preprocess the headlights and determine the thresholds of the three channels. Ref. [40] used a decision tree based on appearance features to enhance the ability to detect headlights. Ref. [25-28] used the data augmentation method of Generative Adversarial Network (GAN) to increase the night training set and strengthen the training of the classifier. Literature [2] adopted an SVM classifier, adding object suggestion of edge box and multi-scale image enhancement technology based on Retinex to improve detection performance. Kuang [12] used local features and image region similarity object proposal methods, which are augmented with learned weights to enhance the reliability of proposals. To improve the detection accuracy of nighttime road objects, Shao et al. [16] developed a multi-scale feature fusion algorithm and implemented it using the FteGanOd framework. Their approach involves enhancing the contrast between the vehicle and the background and suppressing the influence of

ambient light. In particular, the proposed method leverages the FteGanOd technique to facilitate the integration of multi-scale features, thus enabling more precise and comprehensive feature extraction. Therefore, deep learning-based methods are more versatile and robust. Existing nighttime target detection methods based on visual images focus on vehicle light features to detect vehicles, while ignoring the presence of pedestrians and other objects in the lane. The object detection method based on deep learning is a dual task of object positioning and object recognition, which can be divided into one or two-level detection methods.

A. One-Stage Detectors

For stand-alone detection, the prediction of bounding box and the feed-forward full convolutional network for object classification are combined into a single-stage detector. SSD [14] and YOLO [34] are the earliest proposed single-stage detectors. Although the single-stage detector breaks through the detection speed bottleneck of the two-stage detector, the unbalanced background pixels in the image also reduce the detection accuracy. In Ref.[35], the authors tried to improve the loss function of SSD to improve the detection accuracy. In addition, a lightweight backbone network is used in the architecture of MobileNets [36] to pursue faster computing speed.

YOLO detector has a faster reasoning speed, but the detection accuracy is negatively affected. The YOLO detector predicts the boundary frame and probability of the image region through an anchor based alternative. FCOS [37] does not use the pre-designed anchor frame and constructs the endpoint distance of the bounding box through the object center point, which simplifies the object detection problem. However, due to complex post-processing, the speed is very slow. These detection networks usually have a good performance on ideal image features during the day but become even less accurate when applied at night. The reason is that the model is not trained with the nightly training set, and the single-stage detector needs to handle the tasks of object proposal and classification at the same time so that the model cannot focus on a single task and lose a lot of accuracy..

B. Two-Stage Detectors

The process of two-level detection is divided into two steps, that is, region proposal and classification. These models use anchor boxes as references, propose several regions of interest (ROIs), and then segment objects in object proposals. RCNN [29] serves as a standard two-stage object detection model. A series of proposals are firstly generated by selective search then these proposals are fed to a CNN for feature extraction to obtain bounding box regression and perform classification. Fast RCNN [13] improves detection speed with improved RCNN. Faster RCNN [30] shares features and improves detection efficiency through two-stage connections. Two-level detection is used by researchers to improve detection speed and accuracy in different ways. Examples include grouped convolution ResNXt [31] and Feature Pyramid Network (FPN) [32]. FPN performs cropping by obtaining ROI features of different scales. SPP-Net [33] removes redundancy in the network through a corresponding mapping relationship from candidate regions to features of the full image. This type of method achieves more accurate detection results but is accompanied by complex calculations and high time costs.

III. Method

Object proposal (candidate box extraction) and classification are considered as a whole in single-stage detection, while in two-stage detection they are considered as two parts in one model. This paper proposes a two-level detection model, and the object proposal and classification tasks are respectively handed over to models with different advantages for these two tasks. Although it will increase the complexity of the algorithm, when the two models focus on a single task and are connected in series, they show excellent performance. Based on the idea that one module only focuses on one task, the algorithm is referred to as OMOT for short.

A. Module 1：Focus on the object proposal box

The single-level algorithm YOLO combines the original two-level algorithm detection process into one, which reduces the redundant operation steps of the original detection and greatly reduces the calculation amount of algorithm detection. The YOLO detection network based on Darknet improves the performance of the algorithm by continuously improving and deepening the backbone network. While Darknet is a user-friendly tool that is amenable to customization and enhancement, incorporating it into the algorithm may lead to an increase in the algorithm's size, which can adversely impact the detection efficiency of the network. To address this issue, Module 1 will be improved based on the YOLO architecture.

B. Design lightweight backbone networks

This paper replaces the backbone network of the Darknet with different lightweight networks to test the detection efficiency of the algorithm. The purpose is to design a lightweight backbone network suitable for use under the YOLOv4 framework. The MobileNet lightweight network adopts a method of first expanding the channel and then compressing it, which reduces the number of layers of the network. Then, MobileNetV2 added a lightweight inverted residual block. MobileNetV3 added the SE module additionally. In order to solve the computation problem, the backbone networks of MobileNet and YOLOv4 are improved to make the algorithm lightweight. According to the experimental results, this paper selects MobileNetV2 shown in Fig. 2 as the backbone network. Its function is similar to CSPMarket-53 in YOLOV4.

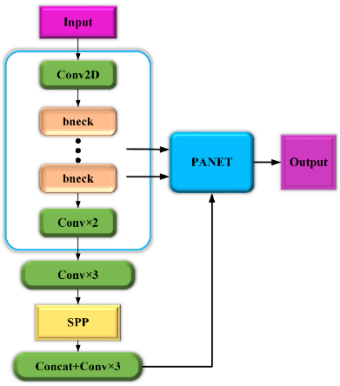

Fig. 2. YOLOv4 based on MobileNetV2

The original YOLOv4 model performs SPP and PAN feature fusion on the Feature map of 8, 16, and 32 times of downsampling. Therefore, it needs to be improved in the corresponding three-layer feature map. Numerous lightweight backbone networks follow the principle that reducing the tensor dimension of the lower network layers can lead to smaller multiplication calculations in the convolutional layers. By maintaining a low-dimensional tensor throughout the entire network, the overall computational speed can be significantly improved. The advantages of lightweight are reflected. When the filter of the convolutional layer only extracts features for low-dimensional tensors, a lot of information contained in the whole will be ignored. As an Expansion layer, MobileNetV2 can expand dimensions. Deep separable convolution can be used to extract features and Projection layer is used to compress data to make the network smaller again. The Expansion layer and Projection layer are characterized by learnable parameters that enable the network to better learn the expanded data dimensions and subsequently recompress the data. By incorporating these layers, the network can effectively enhance its capacity to capture complex patterns and features in the input data, which can ultimately improve its performance. MobileNetV3 has updated blocks and added attention mechanism SE modules. Compared with MobileNet at the same level, it has the best performance. In this paper, the SE module is replaced by improved SPP module of the original model that no longer requires the SE module. The subsequent ablation experiment will be verified in Section 4.

C. Improved SPP module

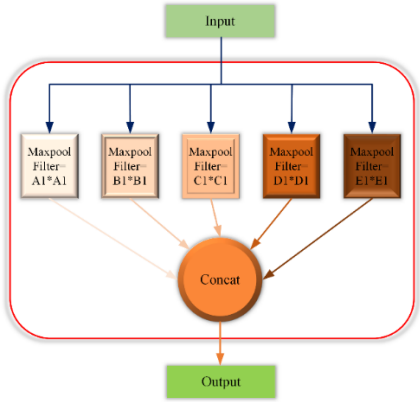

Fig. 3. Improved SPP module diagram

This paper proposes an improved SPP module that can perform a multi-scale fusion of local features and global features, which is of great help to subsequent region proposals. There are only three different feature sizes in the SPP module, which cannot adapt to the road feature extraction under complex conditions such as low visibility at night, so the SPP module needs to be improved. The improved SPP structure is shown in Fig. 3. Two feature sizes are added to adapt to the complex environment of night roads. The feature maps extracted from MobileNetV2 are used to perform maximum pooling operations with five different sizes of convolutions to obtain different information. The feature matrix and finally five feature maps of different sizes will be obtained to output the final feature map in the form of tensor splicing.

The result of the feature extraction of the mobileNetV2 backbone network by the PANet module is obtained by up-sampling and down-sampling, which enhances the extraction of feature information by the feature layer. Through the adaptive pooling method, the feature layer and all feature layers are combined, and when down-sampling, the fusion result is sent to the head for regression. The prediction results of the three feature layers are obtained through the PANet structure. YOLO Head divides the input image into corresponding sizes and obtains the position of the candidate area frame by predicting each preset prior block. Its classification function has been removed and no longer gets classification results for objects. Because there is no need to judge the positive and negative objects of the samples in the obtained object proposal box, only the loss function of the bounding box regression is reserved for the feedback learning of the object proposal. When predicting, add the offsets to get the predicted center position, and then combine the height and width of the obtained prior box to finally get the position of the proposal box. Complete-IoU (CIoU) is represented by Equation (1).

$$CIoU = 1 - IoU + \frac{p^2(b, b^{gt})}{c^2} + av \tag{1}$$

The proposed approach employs various parameters to evaluate the similarity between the predicted and real frames. Specifically, b and b^{gt} represent the center points of the predicted and real frames, respectively, while p denotes the Euclidean distance between these two points. Additionally, c denotes the diagonal distance of the smallest bounding box that can encompass both the predicted and real frames. The aspect ratio similarity, denoted as v, is calculated using Equation (2), while the weight function a is defined by Equation (3). These parameters are utilized to quantify the degree of similarity between the predicted and real frames.

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

$$a = \frac{v}{(1 - IoU) + v} \tag{3}$$

Where w and h represent the width and height of the proposal box, and gt represents the ground truth.

D. Module 2： Focus on object classification

Considering the difficulty in collecting object features at night and a large number of road objects, ResNet50 [40] is selected as the baseline network for improvement. For the object classification task in the complex environment of night roads, the classification network not only needs top-level features with good semantic information to enhance network invariance but also needs detailed features that can distinguish similar objects. This paper improves the residual block structure of ResNet50 to replace the original residual block, and at the same time adds the self-attention mechanism (CBAM) to enhance the recognition of favorable features and reduce the interference of other features, so that the network has rich detailed information to identify the object.

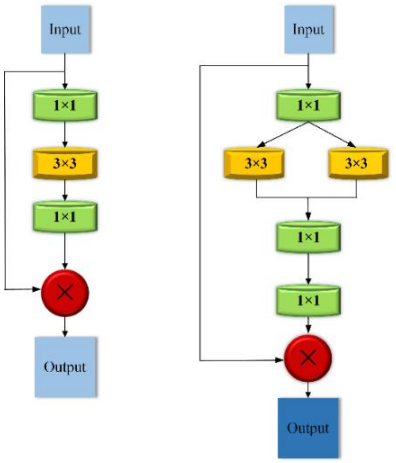

Fig. 4. Improved residual block C1

To speed up the model detection efficiency, this paper proposes an improved residual block. The structure of the residual block is shown in Fig. 4. The feature map input by the previous layer is used for dimensionality reduction by 1×1 convolution. The number of channels of the input feature is expanded by using a bidirectional 3×3 convolution channel to reduce the parameters in the feature layer, to improve the forward reasoning speed of the model. For the image after expanding the number of channels, use a 1×1 convolution to reduce the dimensionality of the forward input, and finally use the 1×1 convolution to perform the dimensionality reduction operation to output the result. After deepening the residual block structure, the network performance will be further improved, the network convergence will be accelerated, and the classification accuracy will be increased.

Add 4 convolutional layers as branches to extract the underlying features, and add CBAM at the same time. The improved network structure is shown in Fig. 5. The input is the result obtained by Module1, to the improved residual block C1, and then enriches the extracted features through CBAM. Branch processing starts later. The first branch uses 4 3×3 convolutions to extract the underlying features of the image, and the second branch uses 3 residual blocks to extract the top-level features for tensor splicing operations. The difference from the original residual block is that the depthwise separable convolution is introduced into the C2, C3, and C4 blocks, and the fully connected layer is improved. The original fully connected layer is changed to a double fully connected layer to further enhance the classification ability of the model.

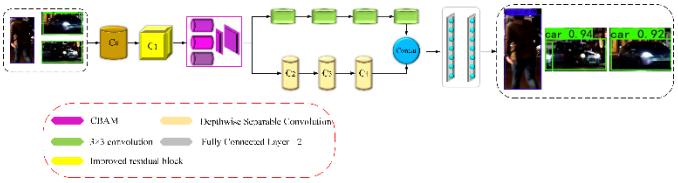

Fig. 5. Network structure diagram of Module 2

The low-level features output by the bottom network can focus more on the small details of the object in the night road environment. Even if there is light or blur interference in the detected image, effective features can be extracted. The high-level features output by the top network tends to observe the important features of the object, can accurately classify the object proposals obtained by Moudle 1, and the improved network can identify the key features of a single object. Bottom-level features and top-level features are very important to improve the refinement classification of objects in fuzzy detection. Therefore, the fusion of bottom-level features and top-level features for refined classification has high accuracy.

IV. EXPERIMENT

All experiments are configured on a running memory of 32GB, the type of server CPU is Intel(R) Core(TM) i5-12400F CPU @2.5GHz2.50GHz, and dual GTX 3090 graphics cards.

A. Night-Time Multiclass Vehicle Dataset

The performance of the proposed network was evaluated using two publicly available datasets. The first dataset used was the Berkeley Deep Drive (BDD) dataset [38], which consists of 100,000 real-world driving scene images. This dataset is known for its large scale, diversity, and complexity of annotations, and is divided into a training set (70,000 images), test set (20,000 images), and validation set (10,000 images). The labels provided for each image include 10 categories, different weather conditions, various types of driving scenes, and the time of day (dawn, dusk, day, and night). The second public dataset is the Hong Kong nighttime multi-level vehicle dataset (HK) [12]. This dataset contains four types of vehicle labels: 1) cars, 2) taxis, 3) buses and 4) vans, plus Upper background (negative sample). The detection set includes 836 sets of images. The night images contained in the BDD dataset are filtered by time tags. And modify the labels of the two datasets, Uniform label definition for all vehicle types: car; retain the label category of the person. The final training dataset is 31,054 images, and the test dataset contains 4672 images. Among them, the test data set is dedicated to two kinds of evaluations for the proposed method, divided into categories to detect 3526 images, 946 images are used to evaluate the object frame. The input images fed into the network have a uniform size of 416x416 pixels and are represented in the RGB color space. It is important to note that

the images used for object detection and those utilized for object proposals are mutually exclusive and have no connection to the training data.

B. Evaluation Metrics

The degree of overlap between the proposed and ground truth bounding boxes is commonly quantified using the Intersection over Union (IOU) metric, which is calculated by dividing the area of intersection between the two boxes by the area of their union [39]. A detected object is considered valid if the IOU between its bounding box and any proposed ground truth exceeds a pre-defined IOU threshold (typically 0.5). The speed of detection is evaluated using Frames per Second (FPS), while Mean Average Precision (mAP) is a widely used metric for assessing the overall performance of object detection algorithms. The Equation (4) of mAP is:

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{4}$$

This experiment has 2 categories, so K is set to 2 and AP is the average precision.

C. Training network

It is carried out under the Ubuntu system, configured in the PyTorch framework, the version of Cuda is 10.2.89, and the version of Cudnn downloaded is 11.2. In order to make the batch size large enough, the batch size is set to 64, and Adaptive Moment Estimation (Adam) is used for optimization. The network starts learning at a rate of 4e-4. The learning rate is gradually decreased with the epoch, divided by 10 every 5 epochs, and the momentum is 0.9. Image enhancement mainly includes random brightness, color dithering, and random contrast. And stipulated that each training iteration is 300 times. A validation phase is performed during training and the best model is saved.

D. Experimental results and discussion

This section presents and discusses the experimental results. As shown in TABLE I, the superiority of the proposed method is verified on different datasets.

TABLE I COMPARISON OF MAP WITH DIFFERENT NETWORKS.

Method	BDD	HK
OMOT(Ours)	0.921	0.936
Improved Resnet50	0.905	0.895
Resnet50	0.843	0.861
MobileNetV2	0.823	0.826
YOLOv4	0. 918	0.908
SSD512	0.852	0.873
Faster RCNN	0.861	0.892

On the two datasets, OMOT obtained superior mAP. Compared with the improved Resnet50 network, OMOT mAP increased by about 2% (BDD) and 4% (HK) after adding Module 1, respectively. Compared with the original Resnet50 network, the improved Resnet50 network has also been greatly improved, and the mAP has increased by about 7% (BDD) and 4% (HK) respectively. Compared with other cutting-edge methods such as MobileNetV2, YOLOv4, SSD512, and Faster RCNN, the proposed method shows superior performance. For the ability of Module 1 to predict the object box, a comparative experiment of IOU is carried out. Usually, the threshold of IOU is set to 0.5, but this leads to all mature models can obtain very high scores. For a clearer comparison, we set the threshold to 0.65, and calculate the IOU of the predicted frame and the rear frame on the verification set as shown in TABLE II.

TABLE II
COMPARISON OF IOU WITH DIFFERENT NETWORKS.

Method	BDD	HK
Moudle1	90.2	91.3
Resnet50	84.3	84.2
MobileNetV2	84.7	81.5
YOLOv4	88.1	89.3
SSD512	79.5	86.3
Faster RCNN	90.4	90.9

The values in TABLE II represent the proportion of each model validation data set greater than the IOU threshold. Because the single-stage detector predicts the object box and classifies it directly, while the two-stage detection is performed separately. So when conducting experiments, we adjusted the labels of the dataset. Only calculates whether the IOU of the predicted bounding box is greater than the set threshold, but does not calculate whether the Confidence Score and the predicted category match the real label. As shown in TABLE II, our Module 1 has the highest proportion of correctly predicted bounding boxes in both datasets.

Fig. 6. Multi-scenario actual measurement map

Fig. 6 shows random examples generated by OMOT, respectively: (a) many people, (b) many vehicles, (c) many people but few vehicles, (d) few people and many vehicles. And split into two lines, representing the output of Module 1 and Module 2. The first row represents the predicted object block diagram output by Module 1, and the second row represents the category recognition of the predicted object block diagram. Fig. 6 shows that OMOT can successfully detect multiple classes in different scenes.

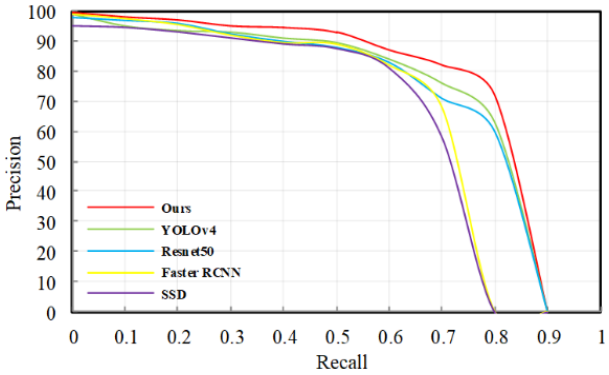

Fig. 7. PR curves of five networks

Fig. 7 shows the PR curves of five networks tested on two datasets. It can be seen that the method proposed in this paper achieves the highest precision under the same recall rate. When the recall rate is 0.5, the precision of OMOT is 93%, which is about 6% higher than the lowest SSD512, and about 3.5% higher than the second-best YOLOv4. And it can be seen from Fig. 7 that the proposed method has always had the best performance as the red curve. Therefore, the method proposed in this paper has higher precision and recall than the other four networks.

TABLE III presents each network's detection speed and parameters with an input batch size of 1 using the same test dataset. Our Module 1 uses YOLOv4-MobileNetV2 to process data faster than the MobileNetV3 backbone, and Module 2 uses the improved Resnet50 to replace the recognition and classification function of YOLOv4-MobileNetV2. Due to the two-level detection, after replacing the improved Resnet50 for classification, the FPS dropped by 4.9 and the processing speed decreased slightly.

TABLE III
COMPUTATIONAL PERFORMANCE.

Method	FPS	Average time(s)
OMOT(Ours)	42.3	0.028
Resnet50	32.0	0.026
SSD512	33.5	0.035
Faster RCNN	12.9	0.092
YOLOv4	41.6	0.024
YOLOv4-MobileNetV2	45.2	0.019
YOLOv4-MobileNetV3	43.4	0.021

Fig. 8 shows the limitations of the proposed method in some scenarios. As shown in Fig. 8(a), long-distance vehicle detections are lost when there are many close-range vehicles. This situation occurs when driving on normal roads, because the current driving conditions of the driving vehicle are mainly considered, so whether it is lane changing or safety issues, long-distance vehicles will not pose a safety hazard. As shown in Fig. 8(b), pedestrians will not be detected when they are not on the road where motor vehicles are driving and are on a dark street corner, which is also an aspect that needs to be improved.

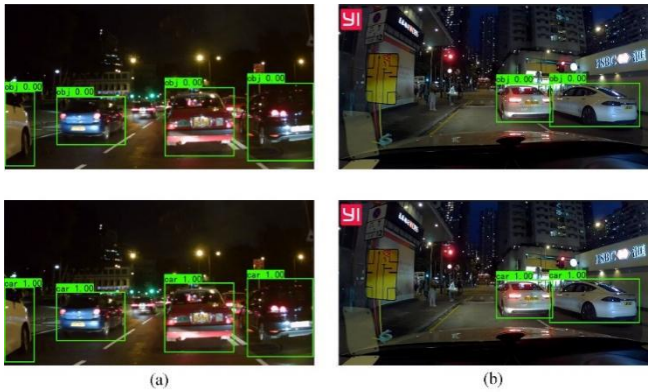

Fig. 8. Example of scene limitations

TABLE IV
ACCURACY COMPARISON OF DIFFERENT NETWORKS UNDER TWO DATASETS

Method	BDD	HK
OMOT(Ours)	92.46%	95.13%
Resnet50	90.63%	92.42%
MobileNetV2	84.70%	89.71%
YOLOv4	91.15%	94.30%
SSD512	89.59%	90.35%
Faster RCNN	90.47%	91.79%

As shown in TABLE IV. In order to ensure the accuracy of the results, each network is tested with the same computing configuration, because the BDD data set is not a traditional night data set, and the label has been modified, so the accuracy has not reached a very good level, but in comparison, the method achieved the highest accuracy of 92.46%. For the HK data set, since this data set only contains the labels of vehicles, and there will be pedestrians on the actual road, it can be regarded as a single-class object detection data set. The proposed method also achieved the highest accuracy rate of 95.13%, which was 0.83% higher than the second place YOLOv4 94.30%.

V. CONCLUSION

This paper proposes an efficient nighttime road object detection method which uses two-stage detection. Module 1 uses YOLOv4 to replace the backbone with MobileNetV2 and modifies the SPP module in the model. Compared with YOLOv4, the number of parameters is reduced, the corresponding total calculation amount is reduced and the detection rate of the object proposal box is improved. Module 2 has made multiple improvements to Resnet50, making Module 2 more sensitive and accurate for object classification tasks. Combine Module 1 with Module 2, each focusing on a single task. Experimental results show that this method effectively enhances the recognition of vehicle features at night, suppresses the interference of other lights, and obtains excellent processing speed and detection ability. Compared with other nighttime vehicle detection methods, our method also specifically detects pedestrians at night, which increases the diversity and

practicability of detection methods and ensures vehicle driving safety.

However, the proposed method still has certain limitations. Some large bus and pedestrians on street corners will be missed. The analysis is due to the fact that the pedestrian objects in long-distance and special scenes contain fewer features and cannot attract the attention of the network. Future work we will continue to improve the network and focus on detection for objects with fewer features.

REFERENCES

[1] Iqbal, Asad, Zia ur Rehman, Shahid Ali, Kaleem Ullah and Usman Ghani. "Road Traffic Accident Analysis and Identification of Black Spot Locations on Highway." *Civil Engineering Journal* 6 (2020): 2448-2456.

[2] H. Kuang et al., "Combining region-of-interest extraction and image enhancement for nighttime vehicle detection," IEEE Intell. Syst., vol. 31,no. 3, pp. 57–65, May/Jun. 2016.

[3] D. Juri´cand S.Lonˇcari´c, "A method for on-road night-time vehicle headlight detection and tracking," in Proc. Int. Conf. Connect. Veh. Expo, 2014, pp. 655–660.

[4] A. López et al., "Nighttime vehicle detection for intelligent headlight control," in Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst., 2008, pp. 113–124.

[5] Y.-L. Chen and C.-Y. Chiang, "Embedded on-road nighttime vehicle detection and tracking system for driver assistance," in Proc. IEEE Int.Conf. Syst. Man Cybern., 2010, pp. 1555–1562.

[6] B. Zoph, D. Cubuk, Ekin, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le,"Learning Data Augmentation Strategies for Object Detection," ArXive-prints, 2019.

[7] Z. Zou and Z. Shi, "Object Detection in 20 Years: A Survey," ArXive-prints, 2019.

[8] Y. Zhang, J. Wang, X. Wang, and J. M. Dolan, "Road-segmentation-based curb detection method for self-driving via a 3d-lidar sensor," IEEE Transactions on Intelligent Transportation Systems, 2018.

[9] I. Gamal, A. Badawy, A. M. Al-Habal, M. E. Adawy, K. K. Khalil, M. A. El-Moursy, and A. Khattab, "A robust, real-time and calibration-free lane departure warning system," in 2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019.

[10]H. Kuang, X. Zhang, Y.-J. Li, L. L. H. Chan, and H. Yan, ''Nighttime vehicle detection based on bio-inspired image enhancement and weighted score-level feature fusion,'' IEEETrans. Intell. Transp. Syst., vol. 18, no. 4, pp. 927–936, Apr. 2017.

[11]L. Chen, X. Hu, T. Xu, H. Kuang, and Q. Li, ''Turn signal detection during nighttime by CNN detector and perceptual hashing tracking,'' IEEE Trans. Intell. Transp. Syst., vol. 18, no. 12, pp. 3303–3314, Dec. 2017.

[12]H. Kuang, L. Chen, L. L. H. Chan, R. C. C. Cheung, and H. Yan, ''Feature selection based on tensor decomposition and object proposal for night-time multiclass vehicle detection,'' IEEE Trans. Syst., Man, Cybern. Syst., vol. 49, no. 1, pp. 71–80, Jan. 2019.

[13]R. Girshick, ''Fast R-CNN,'' in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1440–1448.

[14]Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A. SSD: Single shot multibox detector. In Proceedings of the ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9905 LNCS, pp. 21–37.

[15]Ostankovich, Vladislav, Rauf Yagfarov, Maksim Rassabin and Salimzhan Gafurov. "Application of CycleGAN-based Augmentation for Autonomous Driving at Night." 2020 International Conference Nonlinearity, Information and Robotics (NIR) (2020): 1-5.

[16]Shao, Xiaotao, Caike Wei, Yan Shen and Zhong-li Wang. "Feature Enhancement Based on CycleGAN for Nighttime Vehicle Detection." IEEE Access 9 (2021): 849-859.

[17]C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 391–405.

[18]Kuang, Hulin, Kai-Fu Yang, Long Chen, Yongjie Li, Leanne Lai Hang Chan and Hong Yan. "Bayes Saliency-Based Object Proposal Generator for Nighttime Traffic Images." IEEE Transactions on Intelligent Transportation Systems 19 (2018): 814-825.

[19]Woo, Sanghyun, Jongchan Park, Joon-Young Lee and In-So Kweon. "CBAM: Convolutional Block Attention Module." European Conference on Computer Vision (2018).

[20]J.-M. Guo, C.-H. Hsia, K. Wong, J.-Y. Wu, Y.-T. Wu, and N.-J. Wang, ''Nighttime vehicle lamp detection and tracking with adaptive mask training,'' IEEE Trans. Veh. Technol., vol. 65, no. 6, pp. 4023–4032, Jun. 2016.

[21]X. Dai, D. Liu, L. Yang, and Y. Liu, ''Research on headlight technology of night vehicle intelligent detection based on Hough transform,'' in Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS), Changsha, China, Jan. 2019, pp. 49–52.

[22]J. Chen, J. Chen, and F. Gu, ''Nighttime vehicle detection using deformable parts model,'' in Proc. 7th Int. Conf. Intell. Hum.-Mach. Syst. Cybern., vol. 2, Aug. 2015, pp. 480–483, 2015.

[23]N. Kosaka and G. Ohashi, ''Vision-based nighttime vehicle detection using CenSurE and SVM,'' IEEE Trans. Intell. Transp. Syst., vol. 16, no. 5, pp. 2599–2608, Oct. 2015.

[24]R. O'Malley, E. Jones, and M. Glavin, "Rear-lamp vehicle detection and tracking in low-exposure color video for night conditions," IEEE Trans. Intell. Transp. Syst., vol. 11, no. 2, pp. 453–462, Jun. 2010.

[25]V. F. Arruda, T. M. Paixao, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, ''Cross-domain car detection using unsupervised image-to-image translation: From day to night,'' in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Budapest, Hungary, Jul. 2019, pp. 1–8.

[26]C.-T. Lin, ''Cross domain adaptation for on-road object detection using multimodal structure-consistent image-to-image translation,'' in Proc. IEEE Int. Conf. Image Process. (ICIP), Taipei, Taiwan, Sep. 2019, pp. 3029–3030.

[27]H. Lee, M. Ra, and W.-Y. Kim, ''Nighttime data augmentation using GAN for improving blind-spot detection,'' IEEE Access, vol. 8, pp. 48049–48059, 2020.

[28]C.-T. Lin, S.-W. Huang, Y.-Y. Wu, and S.-H. Lai, ''GAN-based day-to-night image style transfer for nighttime vehicle detection,'' IEEE Trans. Intell. Transp. Syst., early access, Jan. 6, 2020, doi:10.1109/TITS.2019.2961679.

[29]R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, Jun. 2014, pp. 580–587.

[30]S. Ren, K. He, R. B. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in Proc. Neural Inf. Process. Syst., 2015, pp. 91–99.

[31]Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. arXiv 2017, arXiv:1611.05431.

[32]Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.

[33]K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[34]Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.

[35]Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 42, 318–327.

[36]Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv 2017, arXiv:1704.04861.

[37]Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9626–9635.

[38]F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, ''BDD100K: A diverse driving dataset for heteroge-neous multitask learning,'' 2018, arXiv:1805.04687. [Online]. Available:http://arxiv.org/abs/1805.04687.

[39]M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[40]He, Kaiming, X. Zhang, Shaoqing Ren and Jian Sun. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 770-778.

Dear Editor:

I am sending a manuscript entitled “Dual-mode serial night road objectdetection model based on depth separable and self-attention mechanism” authored by qin yang for publication in IEEE Transactions on Vehicular Technology

Existing road detection at night mainly focuses on vehicle detection, but pedestrian detection is also particularly important in the complex environment of night roads. Based on the idea that one module only focuses on one task, this paper proposes a nighttime object detection method under the dual-module framework, which mainly focus on vehicle and pedestrian. In order to perform multi-scale fusion of local features and global features and realize the rapid acquisition of prediction boxes, Module 1 adopts a lightweight depthwise separable network. In Module 2, the recognition of useful features is enhanced by improving the internal structure of the residual block and increasing the self-attention mechanism, thereby improving the network performance and obtaining better classification accuracy. The efficacy of the proposed approach has been evaluated on two benchmark datasets, namely, the Berkeley Deep Drive dataset and the Hong Kong Vehicle dataset. Experimental results demonstrate that the proposed method outperforms several state-of-the-art object detection methods, exhibiting a superior accuracy rate of 93.79% under high standard IoU threshold, while maintaining a high processing speed of 42.3 frames per second (FPS). Consequently, these results verified the potential of the proposed method for advancing the field of object detection in night

For lab reasons, we can only upload partial code containing Moudle 2 improvements with a comparable dataset, please run the requirements file directly for configuration. We obtained a huge performance improvement with a simple classification dataset.

Finally, the authors claim that none of the material in the paper has been published or is under consideration for publication elsewhere. The corresponding author is Dr. Yahong Ma and her address and other information are as follows: Address: School of Information Engineering, Research Center for Internet of Things and Big Data Technology, Xijing University, No.1 Xijing Road, Xi’an, Shannxi 710123, China E-mail: yahongma@sina.com

Thank you very much for your consideration!

Sincerely,
Qin Yang April 1, 2023

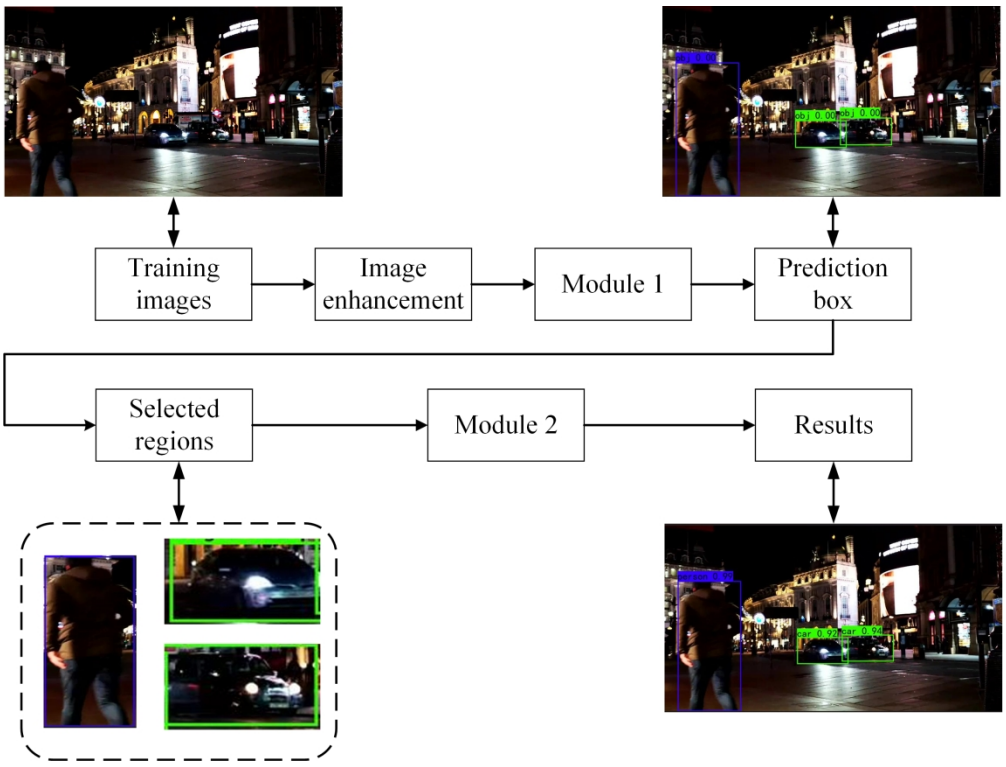

Fig. 1. Flow chart of road object detection at night

236x178mm (300 x 300 DPI)

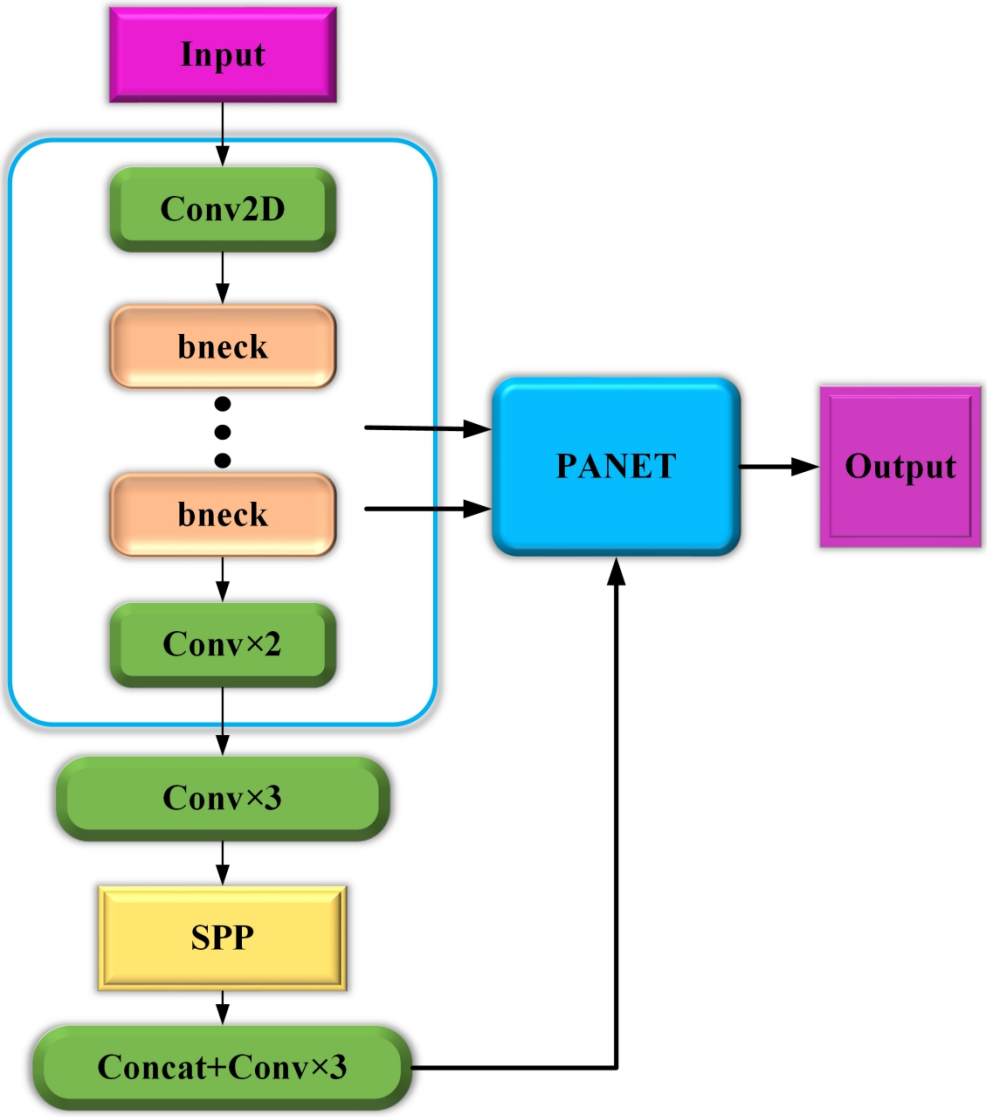

Fig. 2. YOLOv4 based on MobileNetV2

176x198mm (300 x 300 DPI)

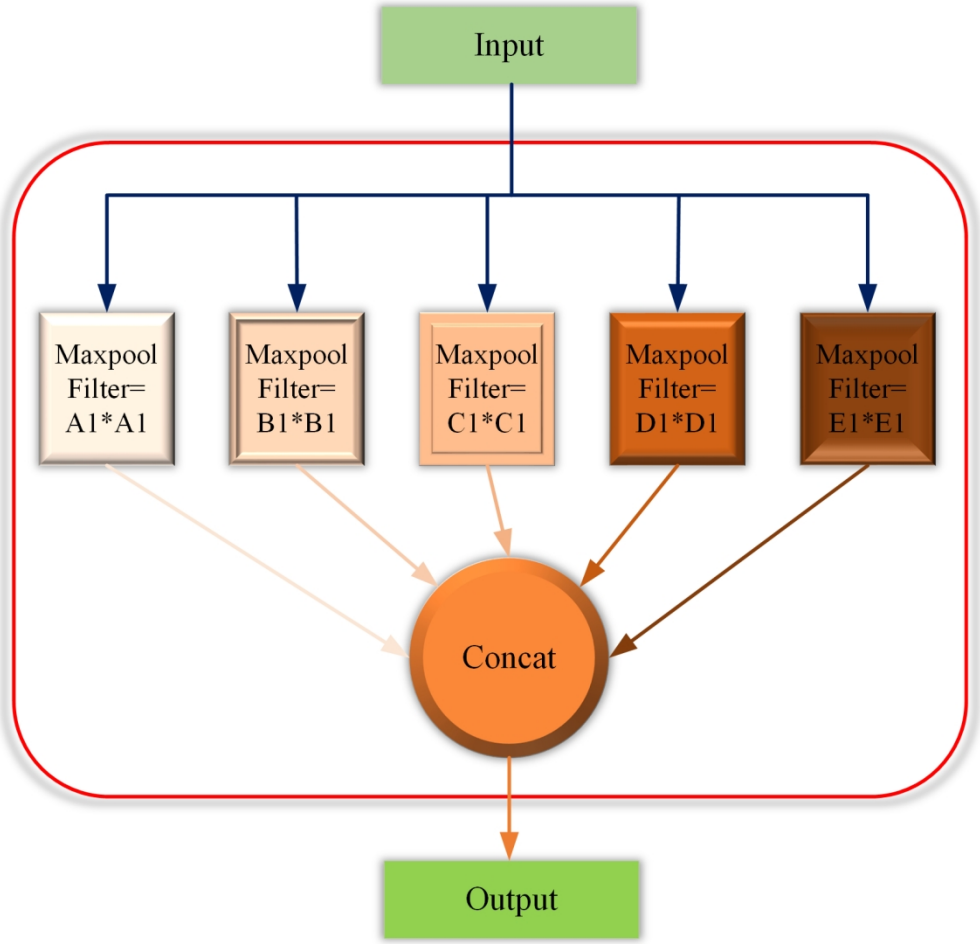

Fig. 3. Improved SPP module diagram

148x141mm (300 x 300 DPI)

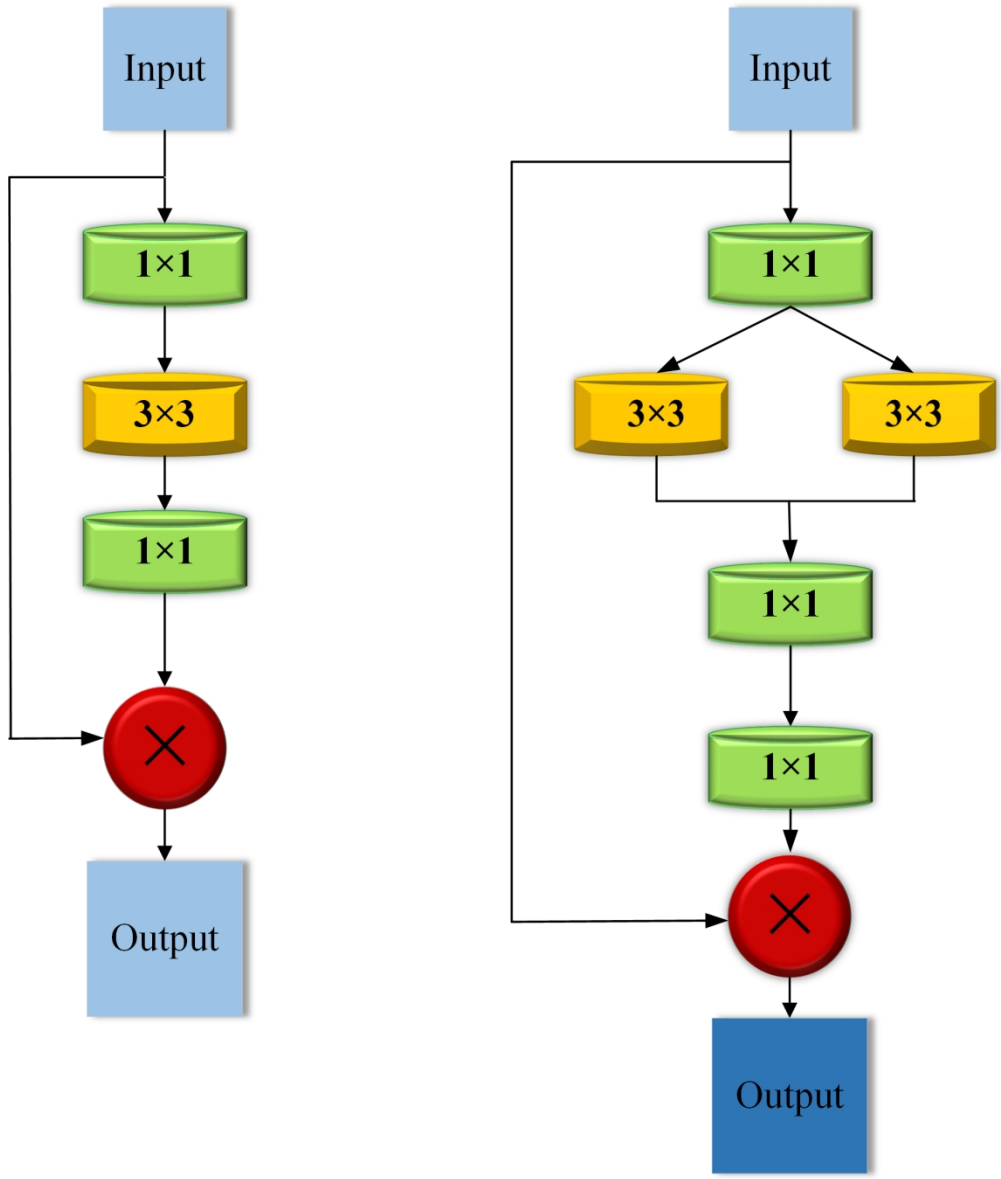

Fig. 4. Improved residual block C1

162x191mm (300 x 300 DPI)

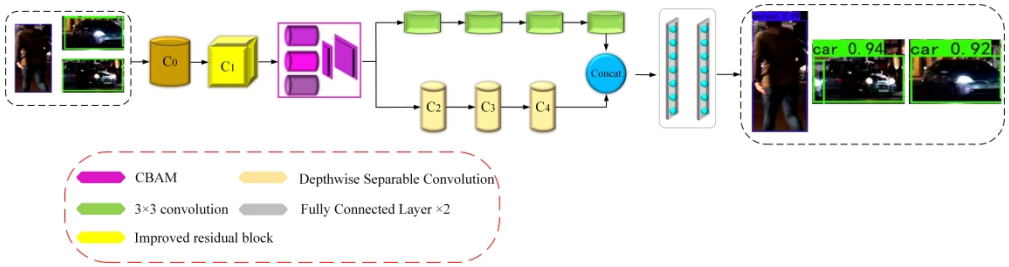

Fig. 5. Network structure diagram of Module 2

594x153mm (300 x 300 DPI)

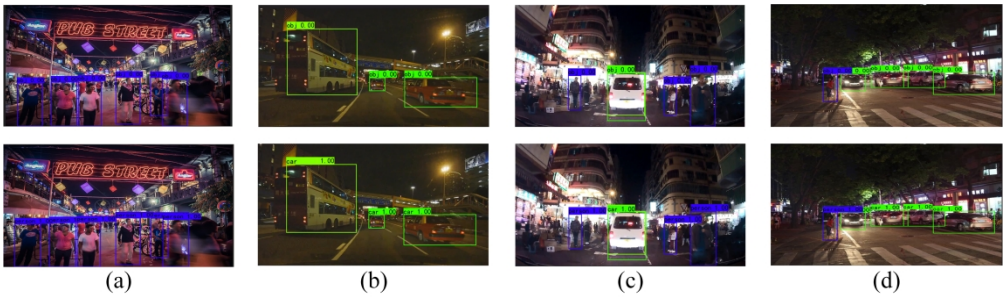

Fig. 6. Multi-scenario actual measurement map

708x213mm (300 x 300 DPI)

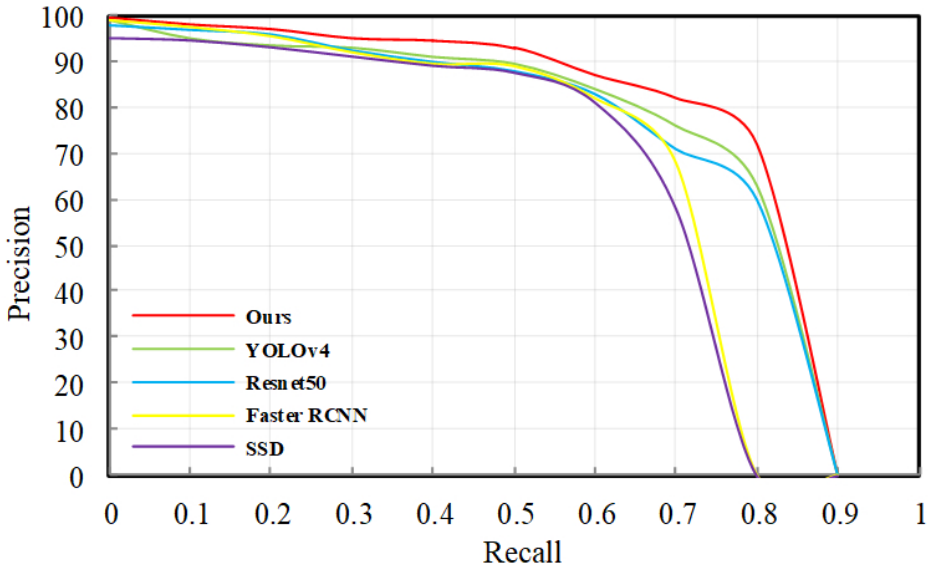

Fig. 7. PR curves of five networks

68x41mm (300 x 300 DPI)

Fig. 8. Example of scene limitations

323x197mm (300 x 300 DPI)

Dual-mode serial night road objectdetection model based on depth separable and self-attention mechanism

Qin Yang, Yahong Ma*, Zeyu Zhao, Linsen Li

***Abstract*—Existing road detection at night mainly focuses on vehicle detection, but pedestrian detection is also particularly important in the complex environment of night roads. Based on the idea that one module only focuses on one task, this paper proposes a nighttime object detection method under the dual-module framework, which mainly focus on vehicle and pedestrian. In order to perform multi-scale fusion of local features and global features and realize the rapid acquisition of prediction boxes, Module 1 adopts a lightweight depthwise separable network. In Module 2, the recognition of useful features is enhanced by improving the internal structure of the residual block and increasing the self-attention mechanism, thereby improving the network performance and obtaining better classification accuracy. The efficacy of the proposed approach has been evaluated on two benchmark datasets, namely, the Berkeley Deep Drive dataset and the Hong Kong Vehicle dataset. Experimental results demonstrate that the proposed method outperforms several state-of-the-art object detection methods, exhibiting a superior accuracy rate of 93.79% under high standard IoU threshold, while maintaining a high processing speed of 42.3 frames per second (FPS). Consequently, these results verified the potential of the proposed method for advancing the field of object detection in night.**

***Index Terms*— Attention Mechanism, Lightweight, Muti-scale Fusion, Night Object Detection**

I. INTRODUCTION

In recent decades, road traffic accidents have only increased, and human errors accounted for as high as 66.8%, which is the main factor affecting road traffic accident. Vehicle errors (25.6%) and environmental factors (7.6%) are respectively secondary, and the third influencing factor [1]. And most of these accidents happen at night. Therefore, nighttime road object detection, as one of the components of advanced assistance systems and autonomous driving, is very important for road safety in intelligent transportation systems (ITS) [2].

Due to the obvious contrast between objects such as vehicles and the background on the daytime road, the features presented by vehicles or other objects are significant. Headlights are currently the feature that most nighttime vehicle detection methods focus on [3-5], because the headlights are turned on at night, and the headlights are noticed as the best feature for nighttime vehicle detection. The task of object detection in autonomous driving presents a host of challenges such as parking and moving vehicles, shadows, adverse weather conditions, lighting variations, and reflections caused by rain and snow, etc. Moreover, in many driving scenes, the headlights of other vehicles are obstructed by surrounding traffic, leading to additional difficulties. Therefore, autonomous vehicles must possess a robust object detection system that can handle diverse driving scenarios, where the detection of headlights is often imprecise. Detailed explanations of object detection techniques are available in the existing literature [6-9].

Kuang et al. [10-12] proposed a method for nighttime image augmentation using bio-inspired techniques, which was used for feature fusion and object classification. In their subsequent work, they combined traditional segmentation and deep learning features to generate regions of interest (ROI) for target detection. This approaches fall under the multi-level learning framework, but it is not an end-to-end learning framework, which can make the training process morecomplex. V et al. [15] used the classical augmentation method and the method based on CycleGAN to improve the accuracy of the night detection perception module, but the appeared visual artifacts had a great impact on the object detection. Shao et al. [16] presented a novel cascaded detection network framework, FteGanOd, that comprises two primary modules, namely, Feature Translation Enhancement (FTE) and Object Detection (OD). The FTE module is responsible for feature translation, whereas the OD module is responsible for detecting objects in the input images. However, there are too many omissions in the detection of long-distance small objects at night and high-speed road conditions, which may cause safety hazards. The current

This work is supported by the General Projects of Shaanxi Science and Technology Plan(No.2023-JC-YB-504), the Shaanxi province innovation capacity support program (No.2018KJXX-095) and the Xijing University special talent research fund (No.XJ17T03) (*Corresponding author: Second Yhong Ma*.) The first author contributed equally.

Yang Qin studied at School of Electronic Information, Xijing University, Xi'an, China, (Email: chaorenyangqin@icloud.com).

Yahong Ma *. Works at at School of Electronic Information, Xijing University, Xi'an, China, (Email: yahongma@sina.com).

Zeyu Zhao works at the School of Computer Science, Xi'an Jiaotong University (email: 1719048836@qq.com)

Linsen Li works at the School of Cyberspace Security at Shanghai Jiao Tong University, is a visiting scholar at the University of California, Los Angeles, and is a member of IEEE (email: lsli@sjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

mainstream deep learning method needs a large amount of data sets for training and the CycleGAN is usually choose to expand the data set. This method can convert styles while maintaining the characteristics of a given input image without collecting more databases. However, there is a bias in the loss function of this method. In a dark environment, the discriminator may give high scores even if the resulting image does not meet the criteria, and the loss function may encourage the generator to generate incorrect behavior.

Through the observation of road, the main objects on the road include vehicles and pedestrians, which can be located anywhere in the image and have different sizes. Therefore, automatic generation of a window set containing various objects is an important step to improve the robustness and accuracy of road object detection. Generate a set of window detection boxes covering all vehicles and pedestrians as data to feed into the classification model. The most common object proposal method named EdgeBoxes [17] consists of a weighted sum of vehicle light detections. In Reference [18], a Bayesian saliency map-based object proposal generator was proposed, which combines multi-scale sliding windows, proposal rejection, scoring, and non-maximum suppression, but also relies too much on vehicle light feature detection and has limitations. Aiming at the above problems, a lightweight convolutional model focusing on object proposals were trained in this paper. The model proposed in this paper does not rely solely on the feature detection of vehicle light and can learn weight adjustment through regional similarity and local contrast in the nighttime road image. The framework of our proposed nighttime road object detection method is shown in Fig. 1.

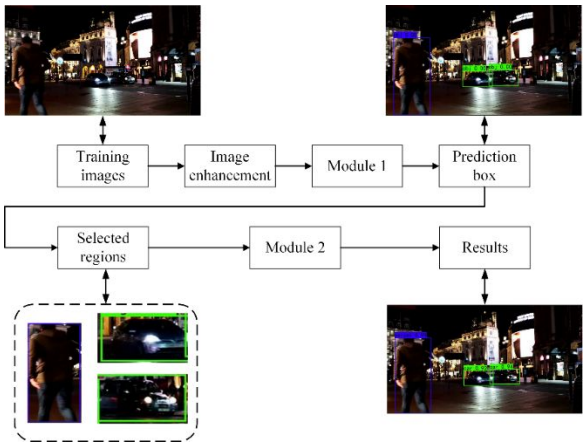

Fig. 1. Flow chart of road object detection at night

There are two modules in the flow chart of road object detection at night, shown in Fig. 1. Module 1 is a lightweight depthwise network, which is used as the object proposal box to improve the efficiency of window detection and reduce the computational burden of the entire detection process. In order to subdivide the extracted night object area Module 2 is modified on the Resnet framework and added the self-attention mechanism CBME[19]. Using the object proposal box predicted by Module 1, only the images in the object proposal box are sent to Module 2 for classification. Through Module 2, feature extraction and category classification are performed on objects in each object frame. The contributions of this paper are outlined below.

1) Aiming at the singleness of car light features in the existing night vehicle detection methods, a lightweight object proposal box model is proposed, which uses the useful features of the object in the image to select the anchor box.

2) A model specially designed for nighttime road object classification is proposed. By extracting the uniqueness of nighttime features in the object proposal box, adding a self-attention mechanism to focus on useful features, and adjusting the model structure according to training feedback to achieve excellent results performance.

3) The dual-mode series of classification after object proposal enables each module to focus on a single task object, increasing detection efficiency and accuracy, and the lightweight network and depthwise separable technology make the entire model more efficient.

The structure of this paper is as follows: Section II outlines the related work that has been conducted in the field of nighttime road object detection. Section III describes the proposed dual-mode serial object detection method in detail, which involves object proposal and classification. The object proposal method and model improvement techniques are presented in this section. The experimental results are discussed in Section IV. Lastly, Section V summarizes the conclusions drawn from this study and highlights future research directions.

II. Related Work

Lights are developed as key information for vehicle object detection at night [3], and headlights and taillights containing red or bright lights are used as regions of interest. Machine learning technology has become the mainstream as nighttime object detection, some classic techniques to obtain object region proposals include threshold segmentation-based methods [20-21], saliency map-based methods [22], or using manually labeled features [23]. Spatial clustering technology with automatic multi-level threshold segmentation is also used to detect vehicles at night [5]. O et al. [24] used the color space of HSV to preprocess the headlights and determine the thresholds of the three channels. Ref. [40] used a decision tree based on appearance features to enhance the ability to detect headlights. Ref. [25-28] used the data augmentation method of Generative Adversarial Network (GAN) to increase the night training set and strengthen the training of the classifier. Literature [2] adopted an SVM classifier, adding object suggestion of edge box and multi-scale image enhancement technology based on Retinex to improve detection performance. Kuang [12] used local features and image region similarity object proposal methods, which are augmented with learned weights to enhance the reliability of proposals. To improve the detection accuracy of nighttime road objects, Shao et al. [16] developed a multi-scale feature fusion algorithm and implemented it using the FteGanOd framework. Their approach involves enhancing the contrast between the

vehicle and the background and suppressing the influence of ambient light. In particular, the proposed method leverages the FteGanOd technique to facilitate the integration of multi-scale features, thus enabling more precise and comprehensive feature extraction. Therefore, deep learning-based methods are more versatile and robust. Existing nighttime target detection methods based on visual images focus on vehicle light features to detect vehicles, while ignoring the presence of pedestrians and other objects in the lane. The object detection method based on deep learning is a dual task of object positioning and object recognition, which can be divided into one or two-level detection methods.

A. One-Stage Detectors

For stand-alone detection, the prediction of bounding box and the feed-forward full convolutional network for object classification are combined into a single-stage detector. SSD [14] and YOLO [34] are the earliest proposed single-stage detectors. Although the single-stage detector breaks through the detection speed bottleneck of the two-stage detector, the unbalanced background pixels in the image also reduce the detection accuracy. In Ref.[35], the authors tried to improve the loss function of SSD to improve the detection accuracy. In addition, a lightweight backbone network is used in the architecture of MobileNets [36] to pursue faster computing speed.

YOLO detector has a faster reasoning speed, but the detection accuracy is negatively affected. The YOLO detector predicts the boundary frame and probability of the image region through an anchor based alternative. FCOS [37] does not use the pre-designed anchor frame and constructs the endpoint distance of the bounding box through the object center point, which simplifies the object detection problem. However, due to complex post-processing, the speed is very slow. These detection networks usually have a good performance on ideal image features during the day but become even less accurate when applied at night. The reason is that the model is not trained with the nightly training set, and the single-stage detector needs to handle the tasks of object proposal and classification at the same time so that the model cannot focus on a single task and lose a lot of accuracy..

B. Two-Stage Detectors

The process of two-level detection is divided into two steps, that is, region proposal and classification. These models use anchor boxes as references, propose several regions of interest (ROIs), and then segment objects in object proposals. RCNN [29] serves as a standard two-stage object detection model. A series of proposals are firstly generated by selective search then these proposals are fed to a CNN for feature extraction to obtain bounding box regression and perform classification. Fast RCNN [13] improves detection speed with improved RCNN. Faster RCNN [30] shares features and improves detection efficiency through two-stage connections. Two-level detection is used by researchers to improve detection speed and accuracy in different ways. Examples include grouped convolution ResNXt [31] and Feature Pyramid Network (FPN) [32]. FPN performs cropping by obtaining ROI features of different scales. SPP-Net [33] removes redundancy in the network through a corresponding mapping relationship from candidate regions to features of the full image. This type of method achieves more accurate detection results but is accompanied by complex calculations and high time costs.

III. Method

Object proposal (candidate box extraction) and classification are considered as a whole in single-stage detection, while in two-stage detection they are considered as two parts in one model. This paper proposes a two-level detection model, and the object proposal and classification tasks are respectively handed over to models with different advantages for these two tasks. Although it will increase the complexity of the algorithm, when the two models focus on a single task and are connected in series, they show excellent performance. Based on the idea that one module only focuses on one task, the algorithm is referred to as OMOT for short.

A. Module 1 ：Focus on the object proposal box

The single-level algorithm YOLO combines the original two-level algorithm detection process into one, which reduces the redundant operation steps of the original detection and greatly reduces the calculation amount of algorithm detection. The YOLO detection network based on Darknet improves the performance of the algorithm by continuously improving and deepening the backbone network. While Darknet is a user-friendly tool that is amenable to customization and enhancement, incorporating it into the algorithm may lead to an increase in the algorithm's size, which can adversely impact the detection efficiency of the network. To address this issue, Module 1 will be improved based on the YOLO architecture.

B. Design lightweight backbone networks

This paper replaces the backbone network of the Darknet with different lightweight networks to test the detection efficiency of the algorithm. The purpose is to design a lightweight backbone network suitable for use under the YOLOv4 framework. The MobileNet lightweight network adopts a method of first expanding the channel and then compressing it, which reduces the number of layers of the network. Then, MobileNetV2 added a lightweight inverted residual block. MobileNetV3 added the SE module additionally. In order to solve the computation problem, the backbone networks of MobileNet and YOLOv4 are improved to make the algorithm lightweight. According to the experimental results, this paper selects MobileNetV2 shown in Fig. 2 as the backbone network. Its function is similar to CSPMarket-53 in YOLOV4.

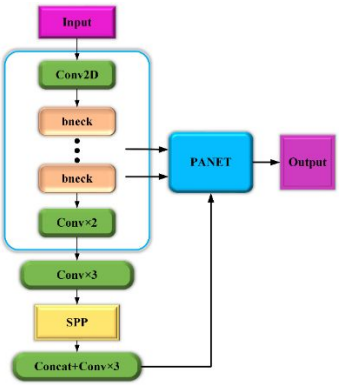

Fig. 2. YOLOv4 based on MobileNetV2

The original YOLOv4 model performs SPP and PAN feature fusion on the Feature map of 8, 16, and 32 times of downsampling. Therefore, it needs to be improved in the corresponding three-layer feature map. Numerous lightweight backbone networks follow the principle that reducing the tensor dimension of the lower network layers can lead to smaller multiplication calculations in the convolutional layers. By maintaining a low-dimensional tensor throughout the entire network, the overall computational speed can be significantly improved. The advantages of lightweight are reflected. When the filter of the convolutional layer only extracts features for low-dimensional tensors, a lot of information contained in the whole will be ignored. As an Expansion layer, MobileNetV2 can expand dimensions. Deep separable convolution can be used to extract features and Projection layer is used to compress data to make the network smaller again. The Expansion layer and Projection layer are characterized by learnable parameters that enable the network to better learn the expanded data dimensions and subsequently recompress the data. By incorporating these layers, the network can effectively enhance its capacity to capture complex patterns and features in the input data, which can ultimately improve its performance. MobileNetV3 has updated blocks and added attention mechanism SE modules. Compared with MobileNet at the same level, it has the best performance. In this paper, the SE module is replaced by improved SPP module of the original model that no longer requires the SE module. The subsequent ablation experiment will be verified in Section 4.

C. Improved SPP module

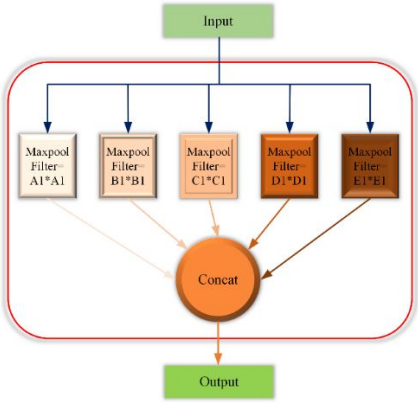

Fig. 3. Improved SPP module diagram

This paper proposes an improved SPP module that can perform a multi-scale fusion of local features and global features, which is of great help to subsequent region proposals. There are only three different feature sizes in the SPP module, which cannot adapt to the road feature extraction under complex conditions such as low visibility at night, so the SPP module needs to be improved. The improved SPP structure is shown in Fig. 3. Two feature sizes are added to adapt to the complex environment of night roads. The feature maps extracted from MobileNetV2 are used to perform maximum pooling operations with five different sizes of convolutions to obtain different information. The feature matrix and finally five feature maps of different sizes will be obtained to output the final feature map in the form of tensor splicing.

The result of the feature extraction of the mobileNetV2 backbone network by the PANet module is obtained by up-sampling and down-sampling, which enhances the extraction of feature information by the feature layer. Through the adaptive pooling method, the feature layer and all feature layers are combined, and when down-sampling, the fusion result is sent to the head for regression. The prediction results of the three feature layers are obtained through the PANet structure. YOLO Head divides the input image into corresponding sizes and obtains the position of the candidate area frame by predicting each preset prior block. Its classification function has been removed and no longer gets classification results for objects. Because there is no need to judge the positive and negative objects of the samples in the obtained object proposal box, only the loss function of the bounding box regression is reserved for the feedback learning of the object proposal. When predicting, add the offsets to get the predicted center position, and then combine the height and width of the obtained prior box to finally get the position of the proposal box. Complete-IoU (CIoU) is represented by Equation (1).

$$CIoU = 1 - IoU + \frac{p^2(b, b^{gt})}{c^2} + av \tag{1}$$

The proposed approach employs various parameters to evaluate the similarity between the predicted and real frames. Specifically, b and b^{gt} represent the center points of the predicted and real frames, respectively, while p denotes the Euclidean distance between these two points. Additionally, c denotes the diagonal distance of the smallest bounding box that can encompass both the predicted and real frames. The aspect ratio similarity, denoted as v, is calculated using Equation (2), while the weight function a is defined by Equation (3). These parameters are utilized to quantify the degree of similarity between the predicted and real frames.

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

$$a = \frac{v}{(1 - IoU) + v} \tag{3}$$

Where w and h represent the width and height of the proposal box, and gt represents the ground truth.

D. Module 2 ：Focus on object classification

Considering the difficulty in collecting object features at night and a large number of road objects, ResNet50 [40] is selected as the baseline network for improvement. For the object classification task in the complex environment of night roads, the classification network not only needs top-level features with good semantic information to enhance network invariance but also needs detailed features that can distinguish similar objects. This paper improves the residual block structure of ResNet50 to replace the original residual block, and at the same time adds the self-attention mechanism (CBAM) to enhance the recognition of favorable features and reduce the interference of other features, so that the network has rich detailed information to identify the object.

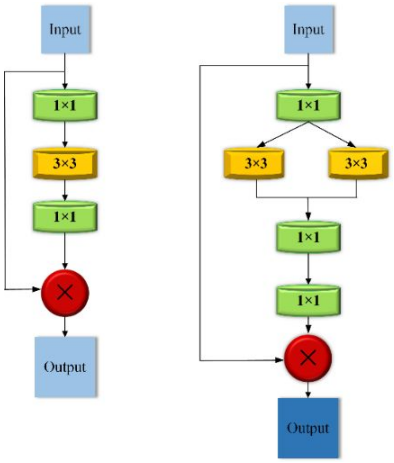

Fig. 4. Improved residual block C1

To speed up the model detection efficiency, this paper proposes an improved residual block. The structure of the residual block is shown in Fig. 4. The feature map input by the previous layer is used for dimensionality reduction by 1×1 convolution. The number of channels of the input feature is expanded by using a bidirectional 3×3 convolution channel to reduce the parameters in the feature layer, to improve the forward reasoning speed of the model. For the image after expanding the number of channels, use a 1×1 convolution to reduce the dimensionality of the forward input, and finally use the 1×1 convolution to perform the dimensionality reduction operation to output the result. After deepening the residual block structure, the network performance will be further improved, the network convergence will be accelerated, and the classification accuracy will be increased.

Add 4 convolutional layers as branches to extract the underlying features, and add CBAM at the same time. The improved network structure is shown in Fig. 5. The input is the result obtained by Module1, to the improved residual block C1, and then enriches the extracted features through CBAM. Branch processing starts later. The first branch uses 4 3×3 convolutions to extract the underlying features of the image, and the second branch uses 3 residual blocks to extract the top-level features for tensor splicing operations. The difference from the original residual block is that the depthwise separable convolution is introduced into the C2, C3, and C4 blocks, and the fully connected layer is improved. The original fully connected layer is changed to a double fully connected layer to further enhance the classification ability of the model.

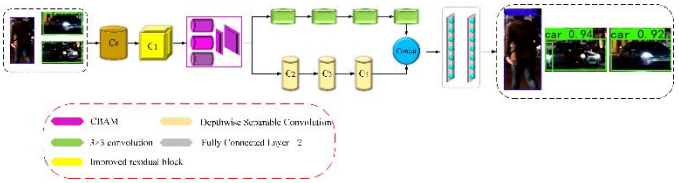

Fig. 5. Network structure diagram of Module 2

The low-level features output by the bottom network can focus more on the small details of the object in the night road environment. Even if there is light or blur interference in the detected image, effective features can be extracted. The high-level features output by the top network tends to observe the important features of the object, can accurately classify the object proposals obtained by Moudle 1, and the improved network can identify the key features of a single object. Bottom-level features and top-level features are very important to improve the refinement classification of objects in fuzzy detection. Therefore, the fusion of bottom-level features and top-level features for refined classification has high accuracy.

IV. EXPERIMENT

All experiments are configured on a running memory of 32GB, the type of server CPU is Intel(R) Core(TM) i5-12400F CPU @2.5GHz2.50GHz, and dual GTX 3090 graphics cards.

A. Night-Time Multiclass Vehicle Dataset

The performance of the proposed network was evaluated using two publicly available datasets. The first dataset used was the Berkeley Deep Drive (BDD) dataset [38], which consists of 100,000 real-world driving scene images. This dataset is known for its large scale, diversity, and complexity of annotations, and is divided into a training set (70,000 images), test set (20,000 images), and validation set (10,000 images). The labels provided for each image include 10 categories, different weather conditions, various types of driving scenes, and the time of day (dawn, dusk, day, and night). The second public dataset is the Hong Kong nighttime multi-level vehicle dataset (HK) [12]. This dataset contains four types of vehicle labels: 1) cars, 2) taxis, 3) buses and 4) vans, plus Upper background (negative sample). The detection set includes 836 sets of images. The night images contained in the BDD dataset are filtered by time tags. And modify the labels of the two datasets, Uniform label definition for all vehicle types: car; retain the label category of the person. The final training dataset is 31,054 images, and the test dataset contains 4672 images. Among them, the test data set is dedicated to two kinds of evaluations for the proposed method, divided into categories to detect 3526 images, 946 images are used to evaluate the object frame. The input images fed into the network have a uniform size of 416x416 pixels

and are represented in the RGB color space. It is important to note that the images used for object detection and those utilized for object proposals are mutually exclusive and have no connection to the training data.

B. Evaluation Metrics

The degree of overlap between the proposed and ground truth bounding boxes is commonly quantified using the Intersection over Union (IOU) metric, which is calculated by dividing the area of intersection between the two boxes by the area of their union [39]. A detected object is considered valid if the IOU between its bounding box and any proposed ground truth exceeds a pre-defined IOU threshold (typically 0.5). The speed of detection is evaluated using Frames per Second (FPS), while Mean Average Precision (mAP) is a widely used metric for assessing the overall performance of object detection algorithms. The Equation (4) of mAP is:

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{4}$$

This experiment has 2 categories, so K is set to 2 and AP is the average precision.

C. Training network

It is carried out under the Ubuntu system, configured in the PyTorch framework, the version of Cuda is 10.2.89, and the version of Cudnn downloaded is 11.2. In order to make the batch size large enough, the batch size is set to 64, and Adaptive Moment Estimation (Adam) is used for optimization. The network starts learning at a rate of 4e-4. The learning rate is gradually decreased with the epoch, divided by 10 every 5 epochs, and the momentum is 0.9. Image enhancement mainly includes random brightness, color dithering, and random contrast. And stipulated that each training iteration is 300 times. A validation phase is performed during training and the best model is saved.

D. Experimental results and discussion

This section presents and discusses the experimental results. As shown in TABLE Ⅰ, the superiority of the proposed method is verified on different datasets.

TABLE I COMPARISON OF MAP WITH DIFFERENT NETWORKS.

Method	BDD	HK
OMOT(Ours)	0.921	0.936
Improved Resnet50	0.905	0.895
Resnet50	0.843	0.861
MobileNetV2	0.823	0.826
YOLOv4	0. 918	0.908
SSD512	0.852	0.873
Faster RCNN	0.861	0.892

On the two datasets, OMOT obtained superior mAP. Compared with the improved Resnet50 network, OMOT mAP increased by about 2% (BDD) and 4% (HK) after adding Module 1, respectively. Compared with the original Resnet50 network, the improved Resnet50 network has also been greatly improved, and the mAP has increased by about 7% (BDD) and 4% (HK) respectively. Compared with other cutting-edge methods such as MobileNetV2, YOLOv4, SSD512, and Faster RCNN, the proposed method shows superior performance. For the ability of Module 1 to predict the object box, a comparative experiment of IOU is carried out. Usually, the threshold of IOU is set to 0.5, but this leads to all mature models can obtain very high scores. For a clearer comparison, we set the threshold to 0.65, and calculate the IOU of the predicted frame and the rear frame on the verification set as shown in TABLE Ⅱ.

TABLE Ⅱ
COMPARISON OF IOU WITH DIFFERENT NETWORKS.

Method	BDD	HK
Moudle1	90.2	91.3
Resnet50	84.3	84.2
MobileNetV2	84.7	81.5
YOLOv4	88.1	89.3
SSD512	79.5	86.3
Faster RCNN	90.4	90.9

The values in TABLE Ⅱ represent the proportion of each model validation data set greater than the IOU threshold. Because the single-stage detector predicts the object box and classifies it directly, while the two-stage detection is performed separately. So when conducting experiments, we adjusted the labels of the dataset. Only calculates whether the IOU of the predicted bounding box is greater than the set threshold, but does not calculate whether the Confidence Score and the predicted category match the real label. As shown in TABLE Ⅱ, our Module 1 has the highest proportion of correctly predicted bounding boxes in both datasets.

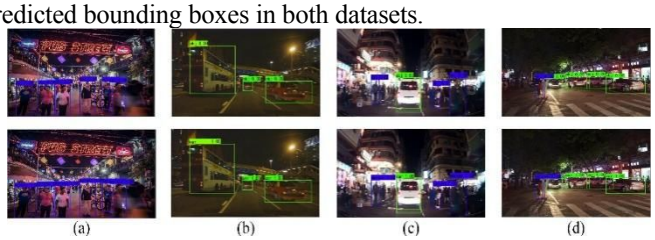

Fig. 6. Multi-scenario actual measurement map

Fig. 6 shows random examples generated by OMOT, respectively: (a) many people, (b) many vehicles, (c) many people but few vehicles, (d) few people and many vehicles. And split into two lines, representing the output of Module 1 and Module 2. The first row represents the predicted object block diagram output by Module 1, and the second row represents the category recognition of the predicted object block diagram. Fig. 6 shows that OMOT can successfully detect multiple classes in different scenes.

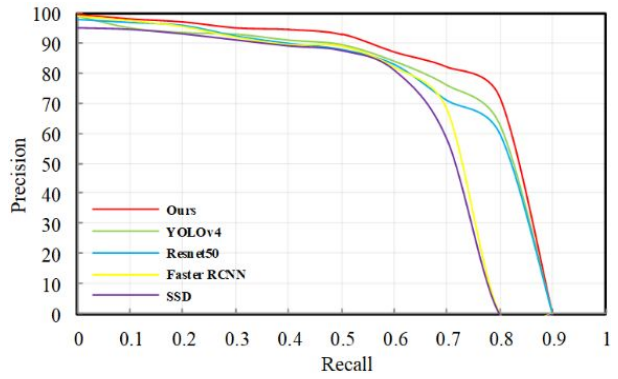

Fig. 7. PR curves of five networks

Fig. 7 shows the PR curves of five networks tested on two datasets. It can be seen that the method proposed in this paper achieves the highest precision under the same recall rate. When the recall rate is 0.5, the precision of OMOT is 93%, which is about 6% higher than the lowest SSD512, and about 3.5% higher than the second-best YOLOv4. And it can be seen from Fig. 7 that the proposed method has always had the best performance as the red curve. Therefore, the method proposed in this paper has higher precision and recall than the other four networks.

TABLE Ⅲ presents each network's detection speed and parameters with an input batch size of 1 using the same test dataset. Our Module 1 uses YOLOv4-MobileNetV2 to process data faster than the MobileNetV3 backbone, and Module 2 uses the improved Resnet50 to replace the recognition and classification function of YOLOv4-MobileNetV2. Due to the two-level detection, after replacing the improved Resnet50 for classification, the FPS dropped by 4.9 and the processing speed decreased slightly.

TABLE Ⅲ
COMPUTATIONAL PERFORMANCE.

Method	FPS	Average time(s)
OMOT(Ours)	42.3	0.028
Resnet50	32.0	0.026
SSD512	33.5	0.035
Faster RCNN	12.9	0.092
YOLOv4	41.6	0.024
YOLOv4-MobileNetV2	45.2	0.019
YOLOv4-MobileNetV3	43.4	0.021

Fig. 8 shows the limitations of the proposed method in some scenarios. As shown in Fig. 8(a), long-distance vehicle detections are lost when there are many close-range vehicles. This situation occurs when driving on normal roads, because the current driving conditions of the driving vehicle are mainly considered, so whether it is lane changing or safety issues, long-distance vehicles will not pose a safety hazard. As shown in Fig. 8(b), pedestrians will not be detected when they are not on the road where motor vehicles are driving and are on a dark street corner, which is also an aspect that needs to be improved.

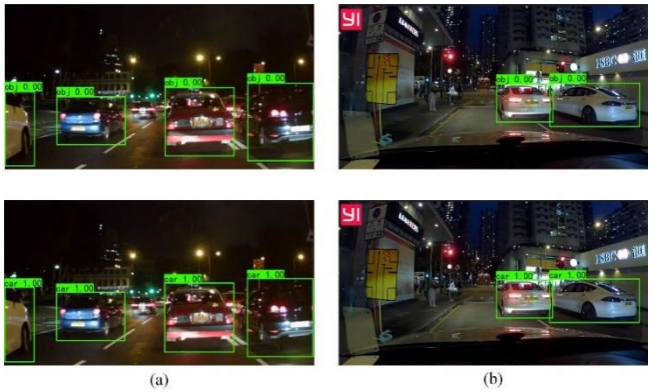

Fig. 8. Example of scene limitations

TABLE Ⅳ
ACCURACY COMPARISON OF DIFFERENT NETWORKS UNDER TWO DATASETS

Method	BDD	HK
OMOT(Ours)	92.46%	95.13%
Resnet50	90.63%	92.42%
MobileNetV2	84.70%	89.71%
YOLOv4	91.15%	94.30%
SSD512	89.59%	90.35%
Faster RCNN	90.47%	91.79%

As shown in TABLE Ⅳ. In order to ensure the accuracy of the results, each network is tested with the same computing configuration, because the BDD data set is not a traditional night data set, and the label has been modified, so the accuracy has not reached a very good level, but in comparison, the method achieved the highest accuracy of 92.46%. For the HK data set, since this data set only contains the labels of vehicles, and there will be pedestrians on the actual road, it can be regarded as a single-class object detection data set. The proposed method also achieved the highest accuracy rate of 95.13%, which was 0.83% higher than the second place YOLOv4 94.30%.

V. CONCLUSION

This paper proposes an efficient nighttime road object detection method which uses two-stage detection. Module 1 uses YOLOv4 to replace the backbone with MobileNetV2 and modifies the SPP module in the model. Compared with YOLOv4, the number of parameters is reduced, the corresponding total calculation amount is reduced and the detection rate of the object proposal box is improved. Module 2 has made multiple improvements to Resnet50, making Module 2 more sensitive and accurate for object classification tasks. Combine Module 1 with Module 2, each focusing on a single task. Experimental results show that this method effectively enhances the recognition of vehicle features at night, suppresses the interference of other lights, and obtains excellent processing speed and detection ability. Compared with other nighttime vehicle detection methods, our method also specifically detects pedestrians at night, which increases the diversity and

practicability of detection methods and ensures vehicle driving safety.

However, the proposed method still has certain limitations. Some large bus and pedestrians on street corners will be missed. The analysis is due to the fact that the pedestrian objects in long-distance and special scenes contain fewer features and cannot attract the attention of the network. Future work we will continue to improve the network and focus on detection for objects with fewer features.

REFERENCES

[1] Iqbal, Asad, Zia ur Rehman, Shahid Ali, Kaleem Ullah and Usman Ghani. "Road Traffic Accident Analysis and Identification of Black Spot Locations on Highway." *Civil Engineering Journal* 6 (2020): 2448-2456.

[2] H. Kuang et al., "Combining region-of-interest extraction and image enhancement for nighttime vehicle detection," IEEE Intell. Syst., vol. 31,no. 3, pp. 57–65, May/Jun. 2016.

[3] D. Juri´cand S.Lonˇcari´c, "A method for on-road night-time vehicle headlight detection and tracking," in Proc. Int. Conf. Connect. Veh. Expo, 2014, pp. 655–660.

[4] A. López et al., "Nighttime vehicle detection for intelligent headlight control," in Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst., 2008, pp. 113–124.

[5] Y.-L. Chen and C.-Y. Chiang, "Embedded on-road nighttime vehicle detection and tracking system for driver assistance," in Proc. IEEE Int.Conf. Syst. Man Cybern., 2010, pp. 1555–1562.

[6] B. Zoph, D. Cubuk, Ekin, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le,"Learning Data Augmentation Strategies for Object Detection," ArXive-prints, 2019.

[7] Z. Zou and Z. Shi, "Object Detection in 20 Years: A Survey," ArXive-prints, 2019.

[8] Y. Zhang, J. Wang, X. Wang, and J. M. Dolan, "Road-segmentation-based curb detection method for self-driving via a 3d-lidar sensor," IEEE Transactions on Intelligent Transportation Systems, 2018.

[9] I. Gamal, A. Badawy, A. M. Al-Habal, M. E. Adawy, K. K. Khalil, M. A. El-Moursy, and A. Khattab, "A robust, real-time and calibration-free lane departure warning system," in 2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019.

[10] H. Kuang, X. Zhang, Y.-J. Li, L. L. H. Chan, and H. Yan, ''Nighttime vehicle detection based on bio-inspired image enhancement and weighted score-level feature fusion,'' IEEETrans. Intell. Transp. Syst., vol. 18, no. 4, pp. 927–936, Apr. 2017.

[11] L. Chen, X. Hu, T. Xu, H. Kuang, and Q. Li, ''Turn signal detection during nighttime by CNN detector and perceptual hashing tracking,'' IEEE Trans. Intell. Transp. Syst., vol. 18, no. 12, pp. 3303–3314, Dec. 2017.

[12] H. Kuang, L. Chen, L. L. H. Chan, R. C. C. Cheung, and H. Yan, ''Feature selection based on tensor decomposition and object proposal for night-time multiclass vehicle detection,'' IEEE Trans. Syst., Man, Cybern. Syst., vol. 49, no. 1, pp. 71–80, Jan. 2019.

[13] R. Girshick, ''Fast R-CNN,'' in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1440–1448.

[14] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A. SSD: Single shot multibox detector. In Proceedings of the ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9905 LNCS, pp. 21–37.

[15] Ostankovich, Vladislav, Rauf Yagfarov, Maksim Rassabin and Salimzhan Gafurov. "Application of CycleGAN-based Augmentation for Autonomous Driving at Night." 2020 International Conference Nonlinearity, Information and Robotics (NIR) (2020): 1-5.

[16] Shao, Xiaotao, Caike Wei, Yan Shen and Zhong-li Wang. "Feature Enhancement Based on CycleGAN for Nighttime Vehicle Detection." IEEE Access 9 (2021): 849-859.

[17] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 391–405.

[18] Kuang, Hulin, Kai-Fu Yang, Long Chen, Yongjie Li, Leanne Lai Hang Chan and Hong Yan. "Bayes Saliency-Based Object Proposal Generator for Nighttime Traffic Images." IEEE Transactions on Intelligent Transportation Systems 19 (2018): 814-825.

[19] Woo, Sanghyun, Jongchan Park, Joon-Young Lee and In-So Kweon. "CBAM: Convolutional Block Attention Module." European Conference on Computer Vision (2018).

[20] J.-M. Guo, C.-H. Hsia, K. Wong, J.-Y. Wu, Y.-T. Wu, and N.-J. Wang, '' Nighttime vehicle lamp detection and tracking with adaptive mask training,'' IEEE Trans. Veh. Technol., vol. 65, no. 6, pp. 4023–4032, Jun. 2016.

[21] X. Dai, D. Liu, L. Yang, and Y. Liu, ''Research on headlight technology of night vehicle intelligent detection based on Hough transform,'' in Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS), Changsha, China, Jan. 2019, pp. 49–52.

[22] J. Chen, J. Chen, and F. Gu, ''Nighttime vehicle detection using deformable parts model,'' in Proc. 7th Int. Conf. Intell. Hum.-Mach. Syst. Cybern., vol. 2, Aug. 2015, pp. 480–483, 2015.

[23] N. Kosaka and G. Ohashi, ''Vision-based nighttime vehicle detection using CenSurE and SVM,'' IEEE Trans. Intell. Transp. Syst., vol. 16, no. 5, pp. 2599–2608, Oct. 2015.

[24] R. O'Malley, E. Jones, and M. Glavin, "Rear-lamp vehicle detection and tracking in low-exposure color video for night conditions," IEEE Trans. Intell. Transp. Syst., vol. 11, no. 2, pp. 453–462, Jun. 2010.

[25] V. F. Arruda, T. M. Paixao, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, ''Cross-domain car detection using unsupervised image-to-image translation: From day to night,'' in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Budapest, Hungary, Jul. 2019, pp. 1–8.

[26] C.-T. Lin, ''Cross domain adaptation for on-road object detection using multimodal structure-consistent image-to-image translation,'' in Proc. IEEE Int. Conf. Image Process. (ICIP), Taipei, Taiwan, Sep. 2019, pp. 3029–3030.

[27] H. Lee, M. Ra, and W.-Y. Kim, ''Nighttime data augmentation using GAN for improving blind-spot detection,'' IEEE Access, vol. 8, pp. 48049–48059, 2020.

[28] C.-T. Lin, S.-W. Huang, Y.-Y. Wu, and S.-H. Lai, ''GAN-based day-to-night image style transfer for nighttime vehicle detection,'' IEEE Trans. Intell. Transp. Syst., early access, Jan. 6, 2020, doi:10.1109/TITS.2019.2961679.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, Jun. 2014, pp. 580–587.

[30] S. Ren, K. He, R. B. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in Proc. Neural Inf. Process. Syst., 2015, pp. 91–99.

[31] Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. arXiv 2017, arXiv:1611.05431.

[32] Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.

[33] K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[34] Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.

[35] Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 42, 318–327.

[36] Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv 2017, arXiv:1704.04861.

[37] Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9626–9635.

[38] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, ''BDD100K: A diverse driving dataset for heteroge-neous multitask learning,'' 2018, arXiv:1805.04687. [Online]. Available:http://arxiv.org/abs/1805.04687.

[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[40]He, Kaiming, X. Zhang, Shaoqing Ren and Jian Sun. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 770-778.