A Review on Object Detection Based on Deep Convolutional Neural Networks for Autonomous Driving

Jialin Lu Shuming Tang Jinqiao Wang Haibing Zhu Yunkuan Wang*

Institute of Automation, Chinese Academy of Sciences, Beijing, 100190

E-mail: lujialin2016@ia.ac.cn

Abstract: Vehicle and pedestrian detection is significant in autonomous driving. It provides information for path planning, lane selection, pedestrian and vehicle tracking, pedestrian behavior prediction, etc. In recent years, the state-of-the-art object detection algorithms have been emerged on the base of deep convolutional neural networks, which can get higher accuracy and efficiency detection results than traditional vision detection algorithms. In this paper, we first introduce and summarize some state-of-the-date object detection algorithms based of deep convolutional neural networks and the improvement ideas of these algorithms. Their frameworks are extracted. Then, we choose several different algorithms and analyze their running results on challenging datasets, Pascal VOC and KITTI. Next, we analyze the current detection challenges as well as their solutions. Finally, we provide insights into use in autonomous driving, such as vehicle and pedestrian detection and driving control.

Key Words: ***Deep Learning; Object Detection;*** **Autonomous Driving*****; Convolutional Neural Networks***

1 INTRODUCTION

Object detection has received great success in recent years. The classical detection algorithms calculate feature maps through a convolution operation on some special feature excitation template and an input image. They get the object location depending on the template's distribution on the input image. Among all classical detection algorithms, Deformable Part Model (DPM) [1] has the highest universality, flexibility, and accuracy. With the rise of deep learning, the CNN-based object detection algorithms are the most popular detection algorithms in object detection challenge instead of classical detection algorithms. They are categorized into two branches: two-stage and one-stage. Both algorithms utilize deep convolutional neural networks to get CNN-feature maps and exploit these feature maps for classification and localization. Recently, some modified algorithms have been processed to improve both detection precision and detection efficiency. The best state-of-the-art CVV-based detection algorithms can reach up to 90% mean average precision (mAP), which is a qualitative leap compared with classical methods. Furthermore, some of them can achieve real-time detection frequency to some extent.

Object detection was applied in face detection at first [2], in which categories and background are relatively simple. With the improvement of detection algorithms, the application scenarios of object detection get more widely, such as, autonomous driving, aviation astronomy, medical imaging, and video surveillance, etc. Object detection also provides significant visual information for other vision tasks.

In this paper, we first summarize the two of the most popular CNN-based object detection branches as well as address the most popular algorithm improvements in region proposal generation, feature extraction subnet, detection subnet, and loss function. Next, we choose several different algorithms and analyze their corresponding results on challenging datasets, Pascal VOC and KITTI. Then, we propose the current detection challenges, such as, partially occlusion, and small object detection problems as well as the solutions. Finally, we provide insights into use in autonomous vehicles, such as vehicle and pedestrian detection and driving control.

2 RELATED WORK

Recent advances in object detection have been deeply and strongly driven by deep convolutional neural networks. The CNN-based object detection algorithms are categorized into two branches: two-stage and one-stage.

The two-stage detection algorithms are region-based convolutional neural network algorithms. The RCNN [3] is the first to exploit the two-stage branch to detect. The detection results of two-stage detection algorithms depend on the CNN-features of the region proposals. The flow chart of the two-stage is illustrated in Figure 1. The two-stage branch has a sweet-spot of high accuracy and relatively fast speed. The one-stage detection algorithms use the CNN-feature maps of the grids to get the final detection results directly. YOLO [4] is the first to exploit the one-stage framework to detect objects. The flow chart of the one-stage is shown in Figure 2. One-stage branch targets on real-time detection and reasonably good accuracy rate.

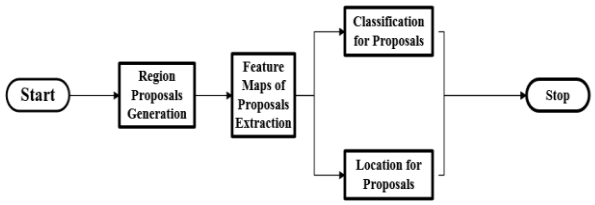

Fig. 1. Flow chart of the two-stage branch

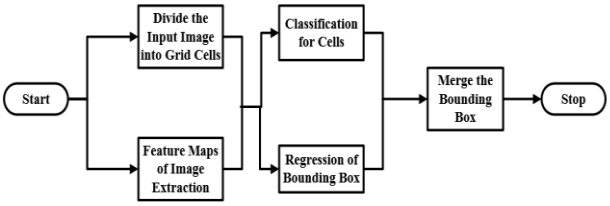

Fig. 2. Flow chart of the one-stage branch

Autonomous driving means vehicle can drive without human's operation after their engines are triggered. It mainly depends on perception and cognition system, monitoring system, decision making system, global positioning system, communication and control systems, etc. The vehicle and pedestrian detection in the traffic scene is significant in autonomous driving. Vehicle detection from in-car videos is critical for the development of autonomous driving systems. It provides information for path planning, lane detection, vehicle tracking. Pedestrian detection aims to complete obstacle avoidance and pedestrian tracking. In 2016, NunoVasconcelos claimed that they had work out a visual detection system, which could detect the pedestrians' behavior at the speed of 2 to 4 frames per second and the error rate is half of other algorithms.

Nevertheless, CNN-based object detection can't get good results in autonomous driving in the real-world at this stage. There are two reasons: One is that vehicles and pedestrians in the traffic scene are always under various poses, scales and occluded by each other or other objects. The other one is that obstacles need to be detected as soon as possible. Therefore, the simple CNN-based object detection algorithms can't apply in autonomous vehicles directly. Over the past decade, a lot of effort had been dedicated to vehicle and pedestrian detection [5, 6].

3 OBJECT DETECTION BASED ON CONVOLUTION NEURAL NETWORKS

There are two popular detection branches in object detection based on the deep convolutional networks. Each has its advantages and disadvantages. In recent years, some improvements have been achieved to modify the algorithms. Some advanced ideas can be utilized in both two-stage and one-stage, such as, feature extraction subnet, detection subnet, etc. Nevertheless, some ideas are aimed at the two-stage, such as region proposals subnet, which is proposed to modify RPN process. While, other ideas are targeted in the one-stage, such as loss function improvement, which is proposed to reduce the imbalance of the samples. These improvements increase the detection precision and the detection efficiency, which will make the object detection based on the deep convolutional networks be utilized in more fields, such as autonomous driving.

3.1 The Basic Framework

1) Two-stage

Ren S, et al. proposed Faster R-CNN[7], which is the milestone of the two-stage object detection algorithm. It can realize end-to-end training and inference with CNN-based features. The detection pipeline of Faster R-CNN is following. Feature extraction networks extract CNN-based features of the input image. Meanwhile, region proposal network (RPN) replaced Search Selective [8] to generate 300 high-quality proposals with more information, which shares the CNN-based features. Then, the proposals with CNN-based features were sent into Fast R-CNN [9] detection subnet. Because of the high-quality proposals and end-to-end training and inference, the speed and accuracy rate of Faster-RCNN can be improved. The framework of Faster-RCNN is shown in Figure 3.

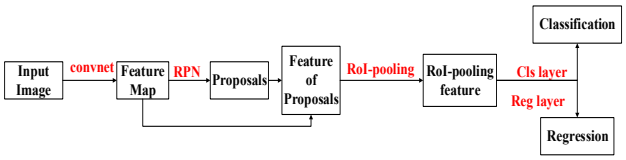

Fig. 3. Framework of Faster-RCNN

2) One-stage

Redmon J, et al. proposed YOLO [4] in 2015, which is the milestone of one-stage. It is worth mentioned that the authors remove the proposals generation item of the two-stage branch and regard the detection problem as a regression problem in YOLO, which can realize the end-to-end training and be extremely fast. In YOLO, an input image is divided into several grids and the whole image is sent into the feature extraction networks to generate features of each grids. Then, several predicted boxes are produced by the features of these grids. Later, these predicted boxes are sent into detection subnet to classify and localize. Finally, Non maximum suppression (NMS) is utilized to eliminate the redundant bounding boxes. In order to improve the detection precision, Liu W, et al. proposed SSD [10] soon after YOLO, in which different resolution features are utilized to detect objects of different scales. High-resolution features are for small objects and low-resolution features are for normal objects' prediction, separately. The framework of the SSD is shown in Figure 4.

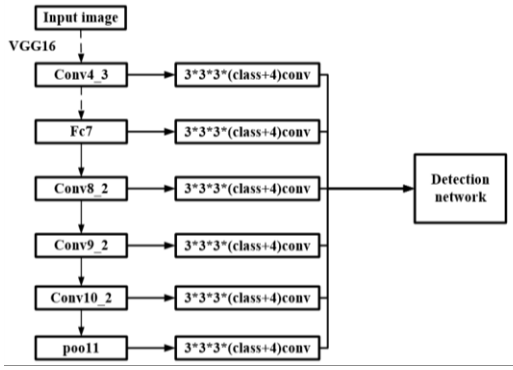

Fig. 4. Framework of SSD

3.2 Improvement both in two branches

1) Feature Extraction Subnet Improvement

Feature extraction subnet is the indispensable part both in two-stage and one-stage branches. They utilize it to extract the CNN-features, which are used to detect the objects. The idea of feature extraction subnet improvement is adding more features, including detailed features and context features. More features can improve the precision of small objects, occlusion objects and truncation objects.

In order to add detailed features, some advanced algorithms [11-14] aim to utilize deep but highly semantic features, intermediate but really complementary features and shallow but naturally high-solution features simultaneously. In the HyperNet [11], the authors utilized max pooling on deep features and compressed them into a uniform space by LPN to generate the Hyper features, which include features of different levels. Hyper features make the proposals have more feature information than Faster-RCNN, which can lead to a higher accuracy rate. The framework of the HyperNet is shown in Figure 5. The idea of top-down pathway and the lateral connection are introduced to build a feature pyramid in FPN [12], which aims to incorporate features of different feature maps. The framework of the FPN is shown in Figure 6. Based on FPN, the researchers add objectness prior into the network in RON [13], thus search space can be reduced. The framework of the RON is shown in Figure 7. Fu C Y, et al. DSSDutilized deconvolution layers to merge feature maps of different resolutions to add detailed features in DSSD [14], which is the similar application of the FPN in one-stage. The part framework of DSSD is shown in Figure 8.

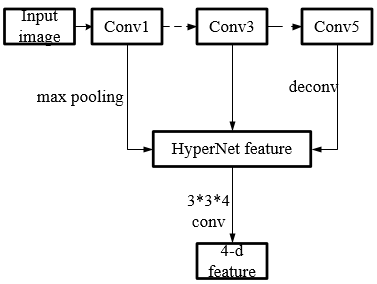

Fig. 5. Framework of HyperNet

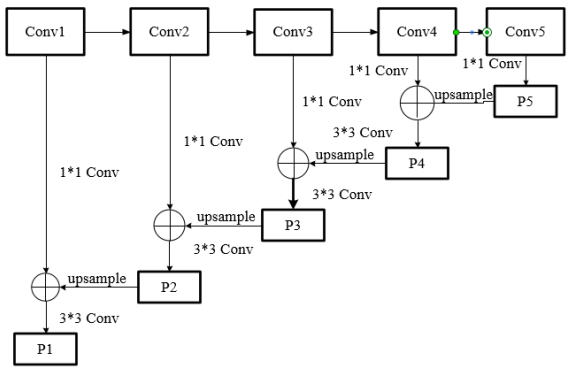

Fig. 6. Framework of FPN

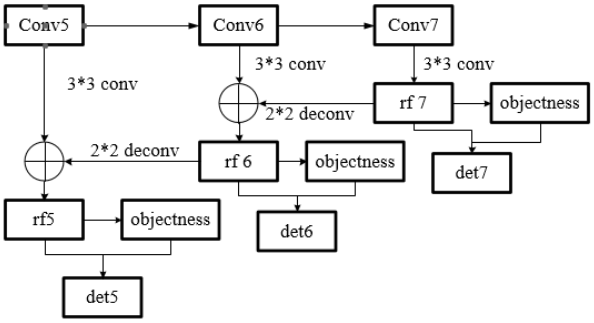

Fig. 7. Framework of RON

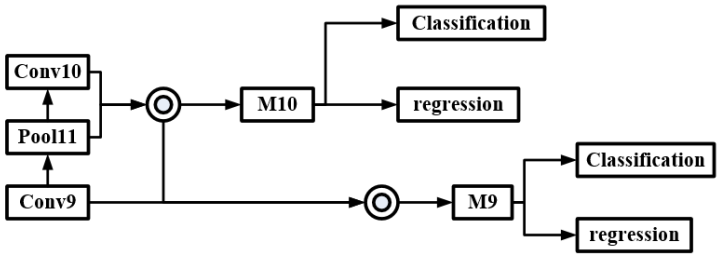

Fig. 8. Part Framework of DSSD

In order to add context features, researchers usually make improved algorithms [15-18] by utilizing some ways to enlarge the ROIs, which will make the ROIs include more context features or use RNN to get context features. The representatives of this improvement are Inside-Outside Net (IONNet) [15], Multi-Scale Deep Convolutional Neural Networks (MS-RCNN) [16], Part and Context Information for Pedestrian Detection with CNNs (PCN) [17], and MDCN[18]. IONNet utilizes two-cascade four-directional IRNN [19] networks to get context feature maps and concats all RoI features to get a global feature map. Two cascade four-directional IRNN [19] networks are utilized to get context feature maps and all RoI features are concated to get a global feature map in IONNet [15]. The idea of MS-RCNN [16] is creating context regions which are 1.5 times larger than object regions. The context regions will add the context features into the origin features. The framework of the MS-RCNN is shown in Figure10. Wang S, et al. used context regions with different scales to extract the context information and maxout for adaptive context selection in PCN [17], which is in order to detect different instances. The part of the framework of PCN is shown in Figure11. The idea of MDCN is similar to PCN. The authors utilize inception filtering units which are produced by the feature extraction network to extract multi-size context features of the original first three levels of the top feature maps. The part framework of MDCN is shown in Figure 12.

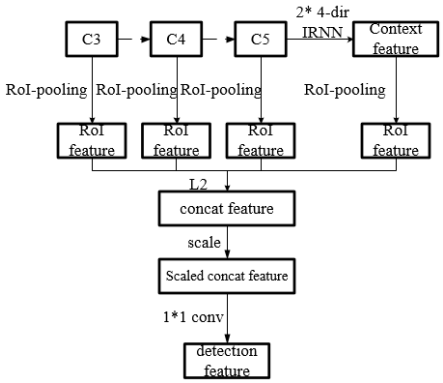

Fig. 9. Framework of IONNet

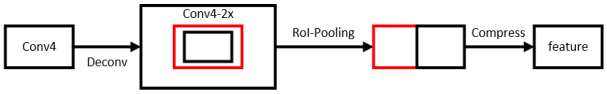

Fig. 10. Part framework of MS-RCNN

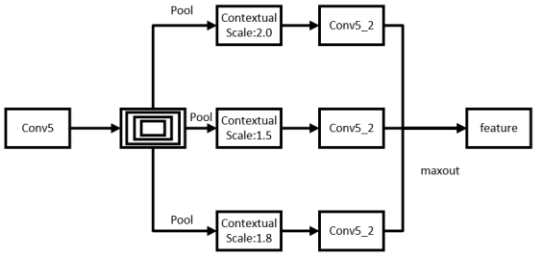

Fig. 11. Part of the framework of PCN

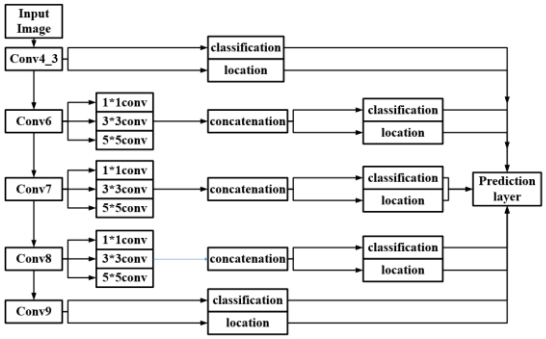

Fig. 12. Framework of MDCN

2) Detection Subnet improvement

Generally, the outputs of the state-of-the-date detection subnets are classification and localization of detected objects. The improvement of detection subnet includes adding other output branch and improving NMS. The former can make the detection subnet improve the detection precision with other information, such as segmentation information. In order to increase the detection precision, another mask branch is produced to do pixel-level classification to increase the accuracy rate in Mask-RCNN [20]. The framework of Mask-RCNN is shown in Figure 13.

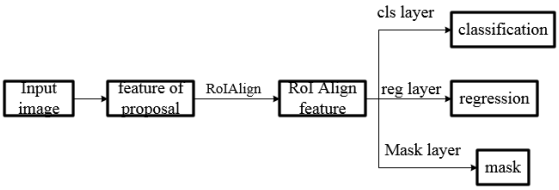

Fig. 13. Framework of Mask-RCNN

For the latter, classical NMS utilizes the classification confidence of the bounding box to eliminate the redundant bounding box of the same instance in the object detection pipeline. There are some shortcomings in such process, such as poor performance in dense object detection and misalignment between classification confidence and localization accuracy. Some advanced NMS algorithms aim to make higher detection precision. The representatives of them are Soft-NMS [21] and IoU-Net [22].

For the classical NMS, when two or more objects of the same class are closed to others, it will regard them as one object and eliminate other objects. In order to make sure that no object is eliminated in the NMS process and increase the precision of dense object detection, Bodla N, et al. propose Soft-NMS to use a decreasing function to decay the detection scores of the bounding box.

Additionally, for the classical NMS, we take classification confidence as the metric to rank the bounding boxes. Problems occur when object with best classification confidence doesn't have the best localization accuracy. To solve this, IoU-Net is utilized to generate the predicted IoU, which replaces the classification confidence to be the ranking keyword in NMS. This will persist the accurately localized bounding box and resolve the misalignment between classification confidence and localization accuracy. The framework of the IoU-Net is shown in Figure 14.

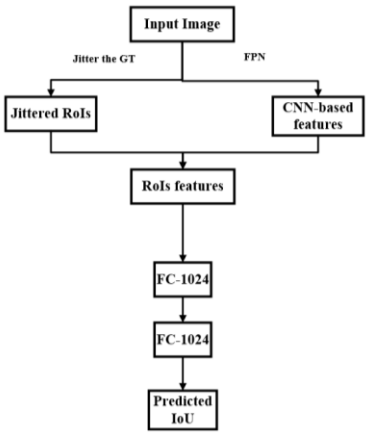

Fig. 14. Framework of IoU-Net

3.3 Special improvement in two-stage

For two-stage branch, the quality and the speed of the proposal generation network (RPN) influence the whole network's detection results directly. Therefore, improving the quality of the proposals can make the detection results better and faster. Zhu Y, et al. introduced a proposal map, named an objectness confidence map, to generate high-quality proposals instead of RPN in Soft Proposal Networks (SPN) [23]. More details can be obtained from

the SP model to improve the precision of classification and localization.

3.4 Special improvement in one-stage

Generally, one major reason of the precision of the one-stage branch is lower than the two-stage branch is that the one-stage detects the anchor boxes directly, which leads to the imbalance of hard and easy samples. To solve the problem, Focal loss function [24] was designed to down-weight the loss assigned to easy examples and up-weights the loss of hard examples to focus on the training of hard ones. Such solving method increases the detection precision.

3.5 Comparisons between Two-stage and One-stage

1) Similarity

Firstly, both of the two branches utilize convolutional neural networks to extract CNN- features, such as, VGGNet, ResNet, and GoogleNet. Secondly, the definition of detection for the two branches is that "detection= (multi) classification + localization". Therefore, the two branches use an image classification framework to classify the objects. They add the localization regression process into the object classification to make the anchor boxes attraction to the ground-truth box. Thirdly, in order to improve the detection precision, the improvements on feature extraction subnet and detection subnet can be used for both two branches.

2) Difference

First of all, the two-stage branch uses some proposals generation methods like RPN to produce proposals, in order to narrow a search space for final detection. This will generate more parameters, which will increase the detection time and dropdown the detection efficiency. On the contrary, the one-stage branch divides the feature maps into several grids and predicts anchor boxes with the center of these grids directly. This will reduce the parameters and increase the speed. Nevertheless, it will cause the imbalance of positive and negative samples. The imbalance leads to worse detection accuracy than the two-stage.

Secondly, although the goals of the two branches are to get high detection precision and detection efficiency, the two-stage actually works better in precision while the one-stage is better at detection efficiency. Obviously, low detection efficiency is the major problem of the two-stage branch. The goal of the efficiency improvement of the two-stage is to reduce the parameters and realize the end-to-end training to increase the speed. For the one-stage, poor detection precision is the major problem. The goal of the detection precision improvement of the one-stage is to solve the imbalance of the positive and negative samples effectively with the improvement of the loss function and so on.

4 THE ANALYSIS OF RESULTS ON CHALLENGING DATASETS

As we know, PASCAL VOC and KITTI are challenging image datasets. PASCAL VOC has many nature images simple environments, which is used for vision tasks mostly. KITTI is often used in real-world traffic scenes with more occlusion or/and small objects. From Refs. [10, 11, 13, 14, 25], we can conclude that: for nature images in simple environments, the best mAP of both two branches are almost the same. Nevertheless, from the detection results in KITTI official website, we can see that for the images in crowd scenes, which need more information to be detected, the best mAP of two-stage is better than one-stage, and one-stage is faster than two-stage for whether simple or crowd environments. Therefore, we could choose different methods for different application scenarios according to requirements of precision and speed.

5 THE CURRENT CHALLENGES AND SOLUTIONS

5.1 Current challenges in object detection

Generally, in the real-world scenarios, detection for small object, occluded object, and dense object remains the most significant challenges.

Small object detection needs more detailed information, which can be obtained from the low level feature maps. Because of the pooling layers and strides, the size of the last feature maps is small and lack of the high-resolution detailed feature information. Therefore, more detailed features should be added to improve the precision. Occluded object detection often occurs in autonomous driving and video surveillance. Objects often gather together and occlude each other in there scenes, which cause the difficulty for detection. Due to the lack of object-body features of occluded objects, the features were extracted by CNN are insufficient. Therefore, add extra features have to be added involved in it. Dense object detection doesn't work well under deep learning, because the NMS exists. Its greedy algorithm makes that some close objects of the same class be regarded as only one object. In order to solve this problem, the NMS process has to be improved in order to make all the proposals can be detected possibly.

5.2 Solution of small object detection

Small object detection needs higher-resolution feature information than normal size object detection. Therefore, we can utilize the improvement of feature extraction subnet to add more detailed information, as mentioned in section 3.2. Nevertheless, such improvement will increase the number of parameters and increase the training and inference time. This will limit the improvement algorithms to be utilized in the real-world scene. Therefore, we can solve the problem if we can simple the network and use the effective information to reduce the number of increasing parameters and the number of the basework parameters.

5.3 Solution of occluded object detection

In general, occlusion is divided into two groups: the inter-class occlusion and the intra-class occlusion. For occlusion, the incomplete of the features lead to the scarcity of information and the crowd occlusions increase the difficulty in object localization. Recent advances in occlusion have been driven by the following methods.

1) Increasing the number of the occlusion examples through data augmentation. It is utilized to generate more hard examples, which can improve robustness of occlusions and deformations detection. Wang X, et al. utilized two subnets, Adversarial Spatial Transformer Network (ASTN) and Adversarial Spatial Dropout Network (ASDN), to generate new occlusion and deformation examples from the original samples in A-Fast-RCNN [25]. ASDN tries to generate a mask on some feature map to get occlusion examples. ASTN, built upon Spatial Transformer Network (STN) [19], produces deformation examples by rotating feature maps. The ASDN and ASTN subnets are cascaded between the ROI-Pooling layer and the full-connected layers and trained with the Fast-RCNN detection subnet jointly. A-Fast-RCNN solves the lack of occlusion and deformation examples in the original datasets and increases the variety of the examples. 2) Adding the context features into original CNN-based features. Because of the incomplete of the features, occlusion object detection can't get enough feature information. Therefore, we can utilize the improvement of the feature extraction subnet to add context features as mentioned in section 3.2. In order to reduce the increasing parameters, we can try to make the anticipation. Because, just occlusion and small object detection need more information we can extract some information from the proposals to get the complexity of them. Thus, we can use different parameters to special object detection, in order to avoid repetitive operation and reduce the parameters.

3) Adding the relation features into original CNN-based features. Relation network in [27] proposed a relation module. The relation module can merge the appearance features generated by the feature extraction network with the geometry features, which are the information of the bounding box and produce the relation features. Relation feature represents the relations of the interaction between the two objects' appearance and geometry. The geometry interaction is lack in the CNN-based features, which are utilized for detection. Therefore, Relation network in [27] added the relation module after either fc layer to transform the 1024-d CNN-based features of all proposals into the 1024-d relation features. Adding relation modules can effectively enhance the detection accuracy. The idea of the relation network comes from the attention mechanism in machine translation [28]. The attention mechanism aims to use the relation between the objects to process a set of objects simultaneously. It will solve some problems caused by detecting object instances individually, which is the disadvantages of all state-of-the-art object detection systems. The framework of the relation module for instance recognition is shown in Figure15.

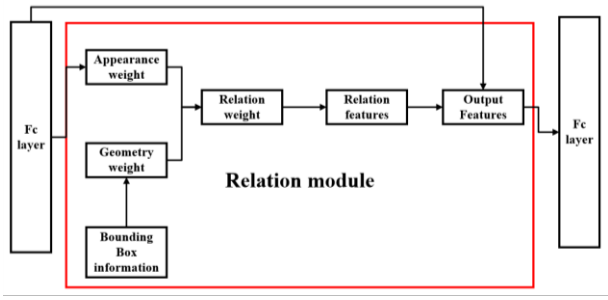

Fig. 15. Framework of the relation module for instance recognition

4) Improving the loss function. Repulsion Loss [29] advanced the loss function proposed in Faster-RCNN. For repulsion loss, the repulsion term is added into original loss to make the bounding box close to its ground-truth box and away from other proposals.

5.4 Solution of dense object detection

Dense object detection is becoming an interesting problem in the computer vision. In the real world, the crowds and the dense vehicles often appear which will increase the difficulty of the object detection. The difference between the dense object detection and occluded object detection is that the dense objects refer to two or more objects are next to each other closely, but occluded objects refer to the part of the occluded objects can't be seen. The problem in dense object detection is that if two or more objects next to each other, the detection result will regard them as a single object after the NMS processing. Therefore, recent advances in dense has been driven by improving the NMS processing.

1) A decreasing function was used in Soft-NMS [21] to decay the detection scores of the bounding box, which will make sure that no object is eliminated in the NMS process. This will reduce the sensitive to the threshold of NMS. In the IoU-Net [22], the network replaced classification confidence with the predicted IoU as the ranking keyword in NMS. This will persist accurately localized bounding box and improve the NMS procedure.

Learning Non-maximum Suppression [30] introduced a new network replacing NMS. It can reduce fully hand-crafted factors in NMS and make the detector less sensitive to the threshold of NMS.

2) We can use relation module [27] to replace the NMS to merge the repeat detecting bounding boxes. Relation module learns the relationship of the two objects' geometric and appearance. The classification score, geometric feature, appearance feature of the object generate the final score for duplicate removal through relation module. The framework of the relation module for duplicate removal network is shown in Figure16. There are three advantages of relation module. First, parameters in relation module can be learned by the training data. This will make the duplicate removal step can be adaptively learnt according to needs, instead of using preset parameters. Second, the relation final score includes the geometric information, which will find the geometric relationships between the close objects. This will improve the dense object detection. Finally, the computation overhead is relatively smaller than the complexity of whole detection networks.

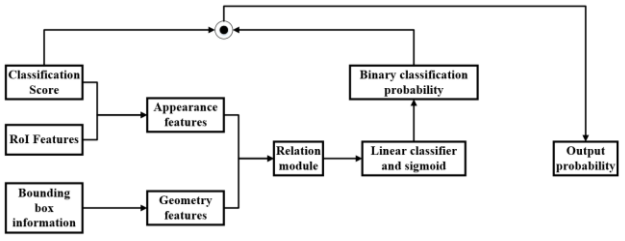

Fig. 16. Framework of the relation module for duplicate removal

Above all, we can see that the improvement of the detection precision always cause the increase of the parameters. This will lead to the low efficiency. The contradiction between

the precision and the efficiency becomes a bad problem in object detection. How to solve this problem will be the first missing in the future.

6 THE APPLICATION IN AUTONOMOUS DRIVING

In autonomous driving, deep convolution networks can be utilized in many aspects, such as vehicle and pedestrian detection, vision localization, and terminal operation control, etc. Vehicle and pedestrian detection is the basis of the driving operation, such as obstacle avoidance, the driving region detection, and path planning. Nowadays, most autonomous driving companies have applied laser radar to complete the object detection in autonomous driving. The advantages of the laser radar are far detection distance and high stability. Nevertheless, laser radar is high cost and has poor performance at object classification. These two shortcomings limit the application of laser radar in autonomous driving, but for cameras, they are no longer a burden. Camera is able to overcome the laser radar's shortcomings. Camera is low cost, which will drive to promote the popularization of the autonomous driving. Nowadays, more and more people have devoted themselves to the algorithm development of computer vision and the technology has been relatively mature. Nevertheless, camera has some shortcomings, which make the camera be abandoned by the autonomous driving company. There are two mainly reasons. One is the high limit of environment and illumination. The other one is the contradiction of the high precision and high speed, which are the indispensable performances in autonomous driving. Therefore, there remains much room for improvement of vision object detection in the future.

Because of the real-time command communication and complex scenes, object detection in autonomous driving needs high speed about 10FPS and high precision for complex object detection in the complex scenes. As we talk before, we can utilize some methods to improve the detection precision, but it will lead the increase of the training time and inference time, which is the worst problem in the object detection. Precision and efficiency have some inverse relationship. Mostly, one is up and the other is down. Therefore, we need to find increase the speed and precision simultaneously. If the speed is too slow, the high-precision detection model can't be used in autonomous driving. Because, when the detection model detects a car in front of you before you crash it, vehicle accidents have already happened and nothing seems to work. Therefore, how to produce a detection model with both high precision and high speed is the significant problem in autonomous driving.

First of all, because of the limited hardware resources and the demand of the real-time detection in autonomous driving, we need a light weight and low latency feature extraction network model. In recent years, there has been rising interest in building small and efficient neural networks. These networks could be utilized in autonomous driving, in order to improve the speed. The representatives of the networks are MobileNet [31], ShuffleNet [32], and SqueezeNet[33]. MobileNet replaces the standard convolutions with the depthwise separable convolutions to generate a new feature extraction network. The depthwise separable convolutions factorize a standard convolution into a depthwise convolution and a 1*1 convolution. The depthwise convolutions apply a single filter to each input channel and a 1*1 convolution kernel to combine the outputs with the depthwise convolution. The structure of the MobileNet is proposed to reduce the computation and model size. MobileNet uses 3*3 depth separable convolution, which is 8 to 9 times less computationally expensive than standard convolution, but with less precision degradation. ShuffleNet is similar to MobileNet. It uses group convolution and channel shuffle operation to make all components in ShuffleNet unit can be computed efficiently. Group convolution helps to reduce the computation complexity. Nevertheless, the stack of the group convolution will block information flow between channel groups and weaken representation. Channel shuffle operation makes it possible to build more powerful structures with multiple group convolutional layers. SqueezeNet proposes a fire model to compress the model. The fire model includes the squeeze convolution layer and the expand layer. As we see that the main improvements of the speed are compression model and reducing parameters. This will also be a learning problem in the future.

The difficulties of the object detection in autonomous driving are vehicle detection and pedestrian detection. Because the environment in real-world is complex and the emergency situation often occurs, which will make the detection more difficult. Therefore, improving the vehicle and pedestrian detection becomes a significant problem in autonomous driving.

Vehicle detection from in-car videos is critical for the development of autonomous driving systems and vehicle detection from surveillance videos is fundamental for the implementation of intelligent traffic management systems. Nevertheless, the existing object detection algorithms are sensitive to the object scales. In the real traffic screen, vehicles gave a large variance of scales, which results a bad detection results. Several algorithms aim to solve this problem. A fast model was presented to detected vehicles with a large variance of scales in SUNet [5]. It includes a context-aware RoI pooling to maintain the contextual and original structure and a multi-branch decision network to detection different scale vehicle through the height of the vehicle.

Pedestrian detection is as a part of occlusion object detection. Although great progress in object detection has been made, pedestrian detection remains a challenging problem due to the diversity of occlusion patterns. Recently, several advances have been driven by loss function improvement, NMS networks improvement, and adding extra feature into feature maps. The literature [6] provided a new network, HyperLeaner, to merge extra feature into feature maps.

Terminal operation control can also be realized with deep convolution networks [34]. The output of the network is the control of steering system brake system. The operation includes turning left, turning right, speeding up and brake. The realization of the terminal operation control advances

the autonomous driving system can be realized by deep convolution network.

7 CONCLUSIONS

Our paper is a review of object detection based on convolutional neural networks. The object detection algorithms are grouped into two branches, two-stage and one-stage. Their representatives are Faster R-CNN and YOLO, respectively. Under the condition of those basic frameworks, the detection performance can be improved by some modifications of feature extraction subnet, detection subnet, and so on. Meanwhile, the detection efficiency can be improved based on narrowing the search space and the quality of the proposals. Through the experiments on challenge datasets, we can see that two-stage is better for object detection in the complex scenes and faster one-stage algorithm runs about 10 times faster than the faster two-stage one. This presents a key issue in the detection of automatic driving targets, subtly balancing the negative correlated detection accuracy and detection efficiency. Therefore, how to propose an algorithm with both high precision and high speed remains a long way, extremely in the real-world scenarios. At the same time, with the rise of deep convolutional neural networks, it can achieve automatic driving systems from more aspects, such as directly driving control. Therefore, we can try our best to propose a model based on deep convolution neural network to finish all the operations in the autonomous driving in the future.

ACKNOWLEDGMENT

This work is supported by the National Key R&D Program of China (No. 2017YFB1002800). We greatly appreciate support from Prof. Qingxiu Du for her instruction and suggestions on this paper. Meanwhile, we also thank for help from all other teachers and students.

REFERENCES

[1] Felzenszwalb P F, Girshick R B, Mcallester D. Cascade object detection with deformable part models[C]// Computer Vision and Pattern Recognition. IEEE, 2010:2241-2248.

[2] Viola P, Jones M. Rapid Object Detection using a Boosted Cascade of Simple Features[C]// Computer Vision and Pattern Recognition, 2001. CVPR 2001.Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2003:I-511-I-518 vol.1.

[3] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2013:580-587.

[4] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. 2015:779-788.

[5] Mao J, Xiao T, Jiang Y, et al. What Can Help Pedestrian Detection?[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:6034-6043.

[6] Hu X, Xu X, Xiao Y, et al. SINet: A Scale-insensitive Convolutional Neural Network for Fast Vehicle Detection[J]. 2018.

[7] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Trans Pattern Anal Mach Intell, 2015, 39(6):1137-1149.

[8] Uijlings J R, Sande K E, Gevers T, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.

[9] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[10] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2015:21-37.

[11] Kong T, Yao A, Chen Y, et al. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection[J]. 2016:845-853.

[12] Lin T Y, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[J]. 2016:936-944.

[13] Kong T, Sun F, Yao A, et al. RON: Reverse Connection with Objectness Prior Networks for Object Detection[J]. 2017.

[14] Fu C Y, Liu W, Ranga A, et al. DSSD: Deconvolutional Single Shot Detector [J]. 2017.

[15] Bell S, Zitnick C L, Bala K, et al. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks[J]. 2015:2874-2883.

[16] Cai, Zhaowei,Fan, Quanfu,Feris, Rogerio S., et al.A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection[J].,2016.

[17] Wang S , Cheng J , Liu H , et al. PCN: Part and Context Information for Pedestrian Detection with CNNs[J]. 2018.

[18] arXiv:1809.01791 [cs.CV]

[19] Le Q V, Jaitly N, Hinton G E. A Simple Way to Initialize Recurrent Networks of Rectified Linear Units[J]. Computer Science, 2015.

[20] He K, Gkioxari G, Dollár P, et al. Mask R-CNN [J]. 2017.

[21] Bodla N , Singh B , Chellappa R , et al. Soft-NMS -- Improving Object Detection With One Line of Code[J]. 2017.

[22] Jiang B , Luo R , Mao J , et al. Acquisition of Localization Confidence for Accurate Object Detection[J]. 2018.

[23] Zhu Y, Zhou Y, Ye Q, et al. Soft Proposal Networks for Weakly Supervised Object Localization[J]. 2017.

[24] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection [J]. 2017:2999-3007.

[25] Wang X, Shrivastava A, Gupta A. A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection[J]. 2017.

[26] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial Transformer Networks [J]. 2015:2017-2025.

[27] Hu H , Gu J , Zhang Z , et al. Relation Networks for Object Detection[J]. 2017.

[28] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. 2017.

[29] Wang X, Xiao T, Jiang Y, et al. Repulsion Loss: Detecting Pedestrians in a Crowd[J]. 2017.

[30] Hosang J, Benenson R, Schiele B. Learning non-maximum suppression[J]. 2017:6469-6477.

[31] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.

[32] Zhang X , Zhou X , Lin M , et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[J]. 2017.

[33] Iandola F N , Han S , Moskewicz M W , et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[J]. 2016.

[34] Xu H , Gao Y , Yu F , et al. End-to-end Learning of Driving Models from Large-scale Video Datasets[J]. 2016.