GJR Publication
OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 05| Issue 02 | March-April | 2025
Journal homepage: https://gjrpublication.com/gjrecs/

Research Article

Effective Target Detection Using Multiple Models of Deep Learning

*Jian Zhu

School of Information Engineering, Yancheng Institute of Technology, Yancheng, China

DOI: 10.5281/zenodo.15310967 **Submission Date: 22 March 2025 | Published Date: 30 April 2025**

Corresponding author:** ***Jian Zhu
School of Information Engineering, Yancheng Institute of Technology, Yancheng, China

Abstract

With the rapid development of intelligent transportation, vehicle and pedestrian target detection techniques are crucial for the development of autonomous driving systems, which can significantly improve road safety. However, in the actual vehicle detection task, target detection for pedestrians and vehicles still faces many challenges. The target detection performance is difficult to meet the requirements when the traffic environment has complex lighting conditions, occlusion of pedestrians and vehicles, varying target motion speeds and high traffic density, as well as insufficient feature information due to the low pixel ratio of the target. Meanwhile, target detection methods need to balance accuracy and real-time performance to ensure their practicality in complex environments. In this paper, we propose MixNet, an effective target detection method based on mixed deep learning models, which realizes multi-scale fast prediction of feature maps at different scales through the lightweight single-stage multi-frame detector MobileNetV2-SSD to capture the location and classification information of the target; and designs the hybrid attention mechanism HAM to fuse the channel-spatial attention to enhance the performance of the target detection method on low-resolution and detail-rich low-level feature maps. The experiments in this paper compare the accuracy and real-time performance under multiple datasets, and the accuracy mAP of MixNet is increased by an average of 12% and the frame rate FPS is improved by an average of 16% compared with DSSD and Faster-RCNN. The method in this paper provides accurate and real-time data support for intelligent transportation system, which is beneficial to the safe and efficient operation of urban transportation and has good practical value.

Keywords: *target detection, multi-scale prediction, feature map, intelligent transportation.*

1. Introduction

With the gradual improvement of urban transportation network and increasing population density, resulting in the urban traffic of pedestrians and vehicles in the intersection and passage of the problem is increasingly prominent, traffic safety hazards increased, seriously affecting the efficiency and safety of traffic. The encounter between pedestrians walking on the road and vehicles is a high incidence of traffic accidents. Especially in congested, complex or low visibility situations, traffic safety is at greater risk [1]. Traditional traffic supervision means relying on manual patrols and traffic signal control are difficult to achieve real-time accurate detection. Advances in intelligent transportation systems based on image recognition and video analytics, especially the rapid development of deep learning technology, can obtain real-time information about the status of pedestrians and vehicles, thus realizing target detection and localization and improving road safety.

Target detection techniques based on different deep learning models can better extract the deep features in the image [2,3] and increase the detection accuracy and robustness, such as convolutional neural network [4], graph neural network [5], generative adversarial network [6] and YOLO family of methods [7]. However, there are different levels of interference in the target background in real environments. Dense areas of road traffic or complex streets confuse the target with the background, and the target is occluded or has too much similarity to background objects. Especially in harsh environments such as low light, nighttime, and haze, it is often difficult for existing detection algorithms to maintain stable accuracy. During the movement of vehicles and pedestrians, the image scale varies greatly and the scales of

pedestrians and vehicles are also very different. At the same time, fast motion can cause the target to be obscured, especially in dense traffic areas where vehicles or pedestrians are partially obscured by other objects. Under the above circumstances, current detection algorithms are prone to the phenomena of missed detection and false detection, especially when the occlusion is serious, it is difficult to effectively detect the occluded target, which seriously affects the detection performance.

In this paper, we propose MixNet, a target detection method based on mixed deep learning models. it utilizes a lightweight single-stage multi-frame detector MobileNetV2-SSD to achieve multi-scale fast prediction and predefined anchor frames to speed up the inference speed of target detection, while designing a hybrid attentional mechanism HAM to better distinguish background and target. In this paper, multiple datasets are used in the experiments and compared with different monitoring networks in terms of both accuracy and real-time performance. The experimental results show that the detection accuracy of MixNet network increases by 12% on average over DSSD[8] and Faster-RCNN[9], and the frame rate FPS is improved by 16% on average.

2. Literature Reviews

Target detection technology is one of the important tasks in the field of computer vision, which is widely used in the fields of automatic driving, intelligent security, medical image analysis, and industrial quality inspection. The core objective of this technology is to accurately recognize the target object from the image or video and accurately calibrate its position, which is mainly classified into non-deep learning approach and deep learning approach.

1. Target Detection Methods for Non-Deep Learning
   The target detection step in this approach is divided into three parts: region selection, feature extraction, and classification [10-13].1) Regions that may contain targets are extracted from the image through the region selection stage, so that the subsequent processing focuses computational resources and processing effort on the key regions; 2) By extracting the image features of a specific region, it provides the necessary information for distinguishing the target from the background; and 3) The classifier Analyzes the extracted features according to a predetermined model to identify object classes in the image.

   The commonly used classifiers for non-deep learning target detection are Support Vector Machine (SVM), AdaBoost, Random Forest, etc. For example, Viola et al. proposed the VJ detector, which utilizes the sliding window technique for target detection, and significantly improves the real-time performance of face detection by combining integral images and cascade classifiers.The HOG algorithm proposed by Dalal et al. generates features by analyzing the gradient histograms of the local regions, and is widely used for pedestrian detection.The deformable part model by Felzenszwalb et al. (DPM), on the other hand, improves the detection accuracy of the model by disassembling the target into multiple parts, which are detected independently and then combined.

2. Deep learning based target detection method
   This approach is able to extract effective features from a large amount of data, which greatly improves the detection accuracy and can handle more complex scenes [4-9]. With the continuous development of deep learning technology, the detection accuracy, speed, and robustness for pedestrians and vehicles are constantly improving. Among them, Convolutional Neural Networks (CNN), YOLO series, Graph Neural Networks (GNN), etc. are the more common methods for modeling target detection networks. AlexNet proposed by Krizhevsky et al. overcame the limitations of the earlier CNNs by introducing a deeper network structure, a large number of parameters and GPU acceleration. Subsequently, the Visual Geometry Group at the University of Oxford designed VGGNet, which used smaller convolutional kernels and increased network depth to significantly improve detection accuracy and achieved excellent performance on multiple datasets. GoogleNet, on the other hand, employs the Inception module, which performs parallel processing based on different features to improve the computational efficiency and expressiveness of the network. Facing the problems of gradient vanishing and training difficulties, Kaiming He et al. constructed ResNet, which utilizes Skip Connection to transfer information more efficiently in the network to avoid gradient vanishing. The development of YOLO series network further improves the performance of target recognition, and improves the detection accuracy and speed through the introduction of ResNet and multi-scale feature extraction, but it is still difficult to meet the complexity of the data sets. and speed, but it is still difficult to meet the detection requirements in complex environments (e.g., low light, bad weather). Current deep learning techniques have made some progress in vehicle and pedestrian target detection, but still need to further improve the performance when dealing with complex backgrounds, scale changes, and target detection in dynamic environments.

3. Target detection network MixNet

For target detection, image feature maps generally include low-level feature maps and high-level feature maps. The former has higher spatial resolution and can capture more detailed information of the target space, while the latter can provide rich contextual and semantic information. Therefore, in this paper, we design MixNet, a target detection network based on mixed deep learning models, by effectively fusing the high-level information of feature maps with the low-level information.

The MixNet network is structured as follows

1. Multi-scale prediction is achieved by a lightweight single-stage oriented multi-frame detector on feature maps of different scales, which represents the semantic information of the input image from low to high level of different feature layers. The detector generates multiple predefined anchor frames for the position of each feature map to capture the location and classification information of the target, and at the same time improves the inference speed of target detection by predicting over the entire image feature map.
2. A hybrid attention mechanism is designed for the "channel─space", which effectively enhances the attention to the important target features of low-level feature maps. The hybrid mechanism can adaptively suppress the influence of irrelevant background, better distinguish between background and target in the case of small targets in the image, and improve the performance of low-resolution and richly-detailed low-level feature maps in the classification task.

3.1 Single-stage multiframe detector

MixNet employs the lightweight single-shot multibox detector MobileNetV2-SSD (MobileNetV2 Single Shot Multibox Detector) for multiscale feature map fusion, which effectively enhances target detection. SSD utilizes a single-stage detection architecture that does not require the generation of candidate regions, thus reducing the computational effort of target detection and generating results in a short period of time. SSD utilizes a single-stage detection architecture that does not require the generation of candidate regions, thus reducing the computational effort of the target detection task and generating detection results in a shorter time, which is suitable for real-time detection tasks. At the same time, the Anchor Boxes mechanism is used to predefine the different scales, aspect ratios, and position information of targets as anchor boxes, and each anchor box corresponds to a candidate region, so that MixNet can directly classify and regress their positions, and can predict the position information of multiple targets in a more efficient way.

3.2 Hybrid attention mechanisms

MixNet employs the Hybrid Attention Mechanism (HAM) to enhance the attention to key features while suppressing irrelevant parts to improve the detection performance of the target. Specifically, the features are dynamically weighted by a self-learning set of weighting coefficients to highlight the attention region to enhance the recognition of key information. HAM includes channel attention and spatial attention.

Channel Attention Mechanism models the correlation between different channels and analyzes the importance of each channel through a network learning method, so as to weight the channels and enhance the features of key channels. It dynamically adjusts the weights between channels according to the contribution of each channel, so that the MixNet network focuses on more important channel features and improves the accuracy of the model in target recognition. First, the input feature maps are subjected to maximum pooling and average pooling, through which different aspects of the spatial information are extracted to obtain two feature maps; then, they are inputted into a two-layer neural network to be processed to obtain two outputs; finally, these two outputs are summed up and normalized by the sigmoid activation function, and the weight coefficients of each channel are obtained for adjusting the feature contribution of each channel that highlights the important features and suppresses the unimportant ones.

Spatial Attention Mechanism focuses on the feature information of a specific location in an image, enabling the network to recognize the location of the target more effectively. It enhances the ability of MixNet network to pay attention to spatial dimensions by weighting features in different spatial regions to enhance the accuracy of target detection and classification. First, the input feature map is processed by maximum pooling and average pooling respectively to obtain two different feature maps; then, the correlation fuses the spatial information so as to obtain the joint feature map; after that, this joint feature map is downscaled and the feature information is simplified to highlight the spatially critical regions; finally, the result is normalized by the activation function to obtain the spatial attention map.

Hybrid Attention Mechanism (HAM) combines channel attention and spatial attention to enhance the effective information of the feature map and suppress unimportant regions. After the above process, the MixNet network strengthens the key information in the low-level feature map and improves the feature extraction capability.

4 Experimentation and Analysis

The experiments in this paper are validated using three different types of datasets, which are CCaltech and KITTI datasets. Elements of these datasets contain targets in different scenarios and are widely used for target detection.

4.1 Assessment of indicators

In this paper, Mean Average Precision (mAP) is used as a comparative index of the accuracy of different network architectures for target recognition monitoring. The specific steps are as follows: firstly, for each category, calculate the precision and recall under different thresholds; secondly, calculate the AP of each category; thirdly, take the average AP of all categories to get mAP.

The inference speed of target detection is a very critical metric in real-time applications. In this paper, we use Frame per Second (FPS) as the speed of different networks, which indicates the number of frames per second that the network model is able to process the image.

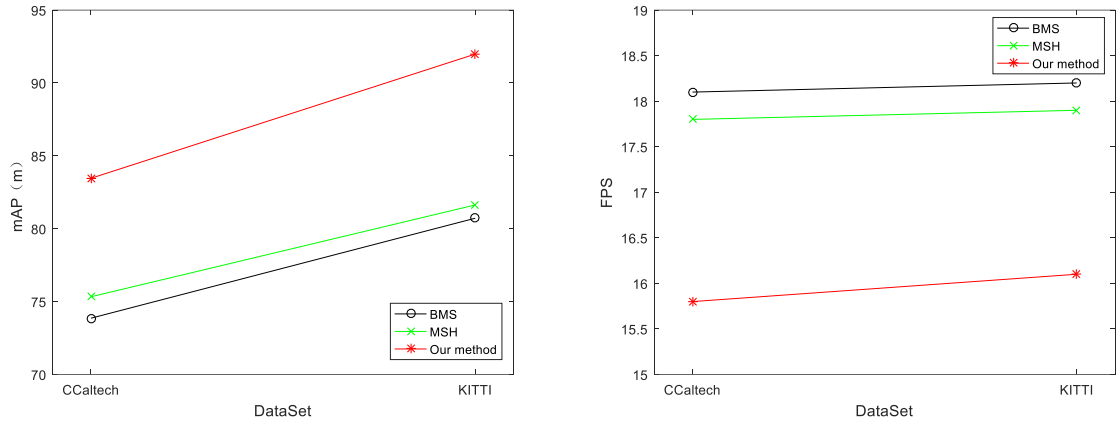
4.2 Experimental Comparison

In this paper, we first perform ablation experiments to compare the performance of module combinations under different scenarios. Subsequently, the MixNet network is compared with deep learning networks such as DSSD and Faster-RCNN(FRCNN) for validation. Finally, the detection results of MixNet network are visualized.

1) ablation experiment

In order to evaluate the performance enhancement of different module combinations, ablation experiments are conducted in this paper, including base MoSSD (BMS), MoSSD in HAM(MSH), and our method. The results are shown Figure 1.

The experimental results show that the MixNet network improves the accuracy of target detection. Its average mAP in the three datasets increases by 14% and 12% over BMS and MSH, respectively. This means that MixNet is able to focus on key regions in the image in low-resolution, occluded, or dense scenes while performing multi-scale feature fusion, which enhances the ability to capture details of the target and thus improves the detection accuracy. In terms of inference speed, the slight decrease in FPS of MixNet is due to the fact that the network introduces more execution steps, thus requiring some increase in computation.



(a) mAP in different combinations

(b) FPS in different combinations

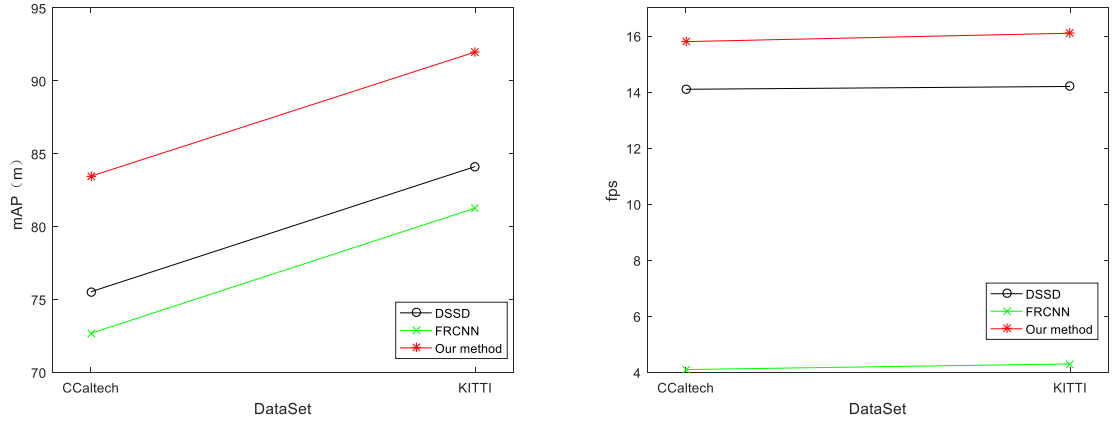
Fig. 1 Different combinations of ablation experiments

Comparing the experimental results on the three datasets, the accuracy of various types of network models is higher on the KITTI and COCO datasets. The reason is that they are mostly static scenes with relatively simple backgrounds and high resolution, so it is easier to extract effective features and obtain relatively accurate detection results.

Overall, the MixNet network is capable of adapting to changes in dynamic environments, especially in the face of small targets and complex backgrounds, with better robustness and accuracy.

2. comparison experiment

The experiments in this section compare the mAP and FPS of DSSD and Faster-RCNN(FRCNN) under different datasets. detection models are shown in Figure 2.



(a) mAP for different methods (b) FPS for different methods

Fig. 5 Comparison experiment of different methods

According to the experimental results, for all three datasets, the target detection accuracy of MixNet network is optimal. For the dataset CCaltech, the mAP of MixNet network is improved by 11% ─ 16% over other methods. For the dataset KITTI, the mAP of MixNet network is improved by 12% and 14%, respectively. In terms of frame rate, the MixNet network also outperforms DSSD and FRCNN with an average improvement of 67%.

Taken together, the MixNet network is able to achieve faster detection speed while improving accuracy. At the same time, it is suitable for labeled detection of multiple different targets in moving/static scenes with better Robustness.

5. Summary

In this paper, we propose a target detection method based on. First, in using MobileNetV2-SSD to achieve multi-scale fast prediction for feature maps of different scales, and use the relationship of the overall features to analyze the important regions in the image. For the irrelevant background information interference in the image, HAM is designed to generate the channel attention area map and spatial attention area map to suppress them effectively, which makes the MixNet network more focused on the feature region and improves the sensitivity and recognition ability of the target. Experimental results show that MixNet network can significantly improve the accuracy of target detection. The work in this paper enhances the efficiency of traffic management and provides more accurate data support for intelligent transportation systems.

References

1. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. CVPR, 2024.
2. Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
3. Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786): 504-507.
4. Gardner M W, Dorling S R. Artificial neural networks (the multilayer perceptron) - areview of applications in the atmospheric sciences[J]. Atmospheric Environment, 1998, 32 (14-15): 2627-2636.
5. XIAO Guoqing, LI Xueqi, CHEN Yuedan, et al. A review of large-scale graph neural network research[J]. Journal of Computing, 2024, 47(01): 148-171.
6. Karras T, Laine S, Aittala M, et al. Analyzing and Improving the Image Quality of StyleGAN[J]. CVPR 2020.
7. Li Y, Ma M, Liu S, et al. YOLO-Drone:A Scale-Aware Detector for Drone Vision[J]. Chinese Journal of Electronics, 2024, 33(04):1034-1045.
8. Yang Zetong, Sun Yanan, Liu Shu, et al. 3dssd: Pointbased 3d single stage object detector[C]// CVPR 2024.
9. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Advances in neural information processing systems, 2015, (1): 91-99.
10. Sabater, Montesano, Murillo. Robust efficient post-processing for video object detection[C]//RSJ International Conference on Intelligent Robots and Systems, 2020.
11. Chin T W, Ding R, Marculescu D. Adascale: Towards real-time video object detection using adaptive scaling[J]. Proceedings of Machine Learning and Systems, 2019.
12. Wu H, Chen Y, Wang N, et al. Sequence level semantics aggregation for video object detection[C] // ICV 2019.
13. Yavariabdi A, Kusetogullari H, Celik T, et al. FastUAV-net: A multi-UAV detection algorithm for embedded platforms[J]. Electronics, 2021, 10(6): 724.

CITATION
Jian Zhu. (2025). Effective Target Detection Using Multiple Models of Deep Learning. In Global Journal of Research in Engineering & Computer Sciences (Vol. 5, Number 2, pp. 124–128).
https://doi.org/10.5281/zenodo.15310967