Proceedings of CONF-CDS 2025 Symposium: Application of Machine Learning in Engineering
DOI: 10.54254/2755-2721/2025.AST26503

HiT-Det: Efficient Target Detection Based on Lightweight Network and Attention Mechanism

Yiwen Chen

Institute of Communication Engineering, Jilin University, Changchun, China
3058105928@qq.com

Abstract. As the main task in computer vision, object detection is aimed at identifying the object in the image and frame its position. Recently, single-stage detection methods, which are on the basis of deep learning (such as YOLO and SSD), have been widely used in automatic driving, monitoring and other fields due to their high detection speed and real-time performance. However, the balance between accuracy and computing resources of existing mainstream detection models still needs to be optimized, especially in resource constrained scenarios. Therefore, this study proposes an efficient target detection method based on lightweight network and attention mechanism. By using lightweight backbone networks (such as MobileNetV2 and Efficient Net lite) in combination with deep separable convolution (DWConv), the amount of parameters and computation of the model is significantly reduced. At the same time, SE attention mechanism (or CBAM, coordinate attention, etc.) is introduced to dynamically adjust the weight of channel and spatial features, so as to enhance target features and restrain background interference. For the sake of optimizing feature mix, we adopt an improved feature pyramid network (FPN) strategy, which enables features of different scales to be fused more efficiently, thus improving the detection ability of small targets. Experimental results show that this method could not only prominently reduce the complexity of the model but also improve the reasoning speed while ensuring the accuracy, especially in the detection of small targets. Compared with YOLOv5s, SSD300 and other classical models, this method perform better in model volume, speed and accuracy. The experiments further demonstrate the effectiveness of the proposed lightweight design and attention mechanism.

Keywords: Target detection, single-stage detection method, separable convolution, lightweight network, attention mechanism

1. Introduction

Target detection is an important task in computer vision [1]. Its aim is to identify multiple targets in a given image, and predict a bounding box and its category for each target. With the rapid advancement of deep learning, the method based on convolutional neural network (CNN) [2] has become a mainstream in the field of target detection. In particular, single-stage target detectors, such as YOLO (You Only Look Once) [3] and SSD (Single Shot Multibox Detector) [4], are widely used in automatic driving, monitoring, unmanned aerial vehicles and other fields due to their high real-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.AST26503

time performance and computing efficiency. YOLO reformulates the object detection task as an end-to-end regression process for prediction, and can quickly detect targets. SSD effectively detects targets of different sizes by designing multi-scale feature maps and anchor frames with various aspect ratios, which further improves the multi-scale detection capability of the model.

Although these methods strike a reasonable balance between accuracy and speed, they still face challenges related to computational resources and model size in practical applications. To meet the needs of resource-constrained environments such as mobile devices and embedded systems, lightweight object detection networks [5] have emerged. In recent years, many studies have reduced the amount of parameters and computation of the model and significantly improved the reasoning speed by using lightweight backbone networks (such as MobileNet [6], ShuffleNet [7]) and deep separable convolution (DWConv) [8] and other technologies. However, lightweight models often involve a trade-off between detection accuracy and speed, particularly when detecting small objects. The traditional lightweight design fails to effectively solve this problem.

To address this issue, this paper proposes an efficient object detection method based on a lightweight network and an attention mechanism. We combine lightweight backbone networks (such as MobileNetV2, Efficient Net lite [9]) with deeply separable convolution, significantly reducing the amount of computation and model parameters. At the same time, by introducing attention mechanism (such as SE, CBAM [10] or coordinate attention), the channel and spatial feature weights in the network are dynamically adjusted, thus improving the expression ability of target features and effectively suppressing the influence of background noise. To further enhance the detection performance for small objects, we design an improved feature pyramid network (FPN) fusion strategy [11] to optimize the integration of multi-scale features.

The innovation of this study lies in: the complexity of the model is greatly reduced by selecting lightweight backbone networks and deeply separable convolutions; The introduction of attention mechanism effectively improves the feature expression of the target; Combined with the improved feature fusion strategy, the detection ability of small targets is enhanced. Compared with YOLOv5s, SSD and other classical models, the results of the experiments demonstrate that the proposed model significantly improves the detection speed while maintaining high accuracy, especially on resource constrained devices.

2. Related work

2.1. Lightweight target detection model

Although traditional detection algorithms such as YOLOv3, Faster R-CNN, etc. have high accuracy, they have large parameters and slow reasoning speed, and are difficult to deploy in resource constrained terminal devices. Therefore, the lightweight detection model came into being. MobileNet series (V1/V2/V3) greatly decreases model parameters and computational workload by introducing depth separable convolution; ShuffleNet uses channel mixing and group convolution to further improve efficiency; GhostNet proposes a mechanism to generate "redundant" features to further compress the network structure. These lightweight backbone networks are widely used in efficient detectors such as YOLOv4 tiny, YOLOv5s, Efficient Det, and have shown good speed precision compromise performance in embedded devices [12].

DOI: 10.54254/2755-2721/2025.AST26503

2.2. Attention mechanism

Attention mechanism effectively enhances the selectivity of feature expression by giving the network the ability to "focus on key areas". Channel attention mechanism, such as the Squeeze and Exception (SE) module, improves the network's responsiveness to key targets by modeling the dependencies between channels [13]; Spatial attention mechanism, such as CBAM (Evolutionary Block Attention Module), combines channel and spatial information for feature reweighting, which enhances the ability of target location and feature suppression; Coordinate Attention further considers the impact of location information on recognition, which is prominent in mobile networks. Embedding these attention modules into the lightweight detection network can effectively improve the robustness and accuracy of the model in complex background, small target detection and other scenarios [14]..

2.3. Second section

Detecting targets at multiple scales is a crucial approach to enhancing the accuracy of small object detection. The feature pyramid network (FPN) combines high-level semantic information with low-level detailed features using a top-down design, thus significantly enhancing the model's perception of objects of various sizes. FPN is particularly effective in solving the problem of multi-scale target detection. Especially in complex scenes, it can effectively improve the detection performance of small targets by combining low-level high-resolution information with high-level semantic information. In order to further enhance the effect of feature fusion, PANet (Path Aggregation Network) is extended on the basis of FPN. PANet introduces a bottom-up path to enhance the expression ability of low-level features, thus enhancing the detection performance of small targets. Through more detailed feature aggregation strategy, PANet improves the robustness of the model in complex scenes, especially for small targets or targets with fuzzy boundaries, its performance is better than that of traditional FPN. In addition, BiFPN (Bidirectional FPN) optimizes the interaction of multi-layer features through weighted fusion and jump connection mechanism. This design not only improves the efficiency of information flow, but also makes the network more efficient in multi-scale feature fusion. BiFPN enables features at different levels to fully interact through two-way feature transmission, and enhances the detection accuracy of targets at different scales. To meet the lightweight computing requirements, researchers also propose a lightweight version of FPN, or use the ghost module instead of the traditional convolution module for feature fusion. Ghost module can not only maintain efficient feature extraction, but also significantly reduce the computational overhead by generating redundant features. These optimization schemes further improve the practicability and flexibility of FPN structure in embedded devices or low computing resource environments, and enhance the balance between speed and accuracy of target detection models [15].

3. Method

This research proposes an efficient target detection method that integrates lightweight network structure and attention mechanism, aiming to achieve lower algorithmic complexity and faster reasoning speed with the detection accuracy, especially for edge computing and embedded application scenarios.

This method mainly includes lightweight backbone network, attention enhancement module, simplified feature fusion structure and unified detection head. The overall architecture is roughly as follows: First, after a series of standardization and preprocessing operations (such as size adjustment

DOI: 10.54254/2755-2721/2025.AST26503

and normalization), the input image is sent to the backbone network to extract basic features. We use MobileNetV2 as the backbone structure to extract three different scale feature maps for detecting large, medium and small targets. Next, all feature maps produced by the backbone network are refined using channel and spatial weights within the Attention Module. This module is composed of SE and CBAM, aiming to improve the perception ability of the model to key areas, especially in the detection of complex backgrounds, small targets or low contrast targets. Then, the attention enhanced multi-scale features are input to the feature fusion module. We use a simplified feature pyramid network (FPN) structure, which combines high-level semantic features with low-level detailed features after upsampling, and finally generates a unified size fusion feature map through convolutional layer fusion. Finally, the fused feature map is sent to the detection head to complete the boundary box regression, confidence prediction and target classification. The detector has a lightweight structure and can process multiple feature maps in parallel, thus ensuring the detection accuracy without significantly increasing the delay. The whole network architecture emphasizes simple structure and efficient combination of modules, integrates attention mechanism and multi-scale feature fusion under the premise of lightweight, and provides strong support for deployment in mobile terminals and real-time scenarios. Figure 1 is the structure diagram.

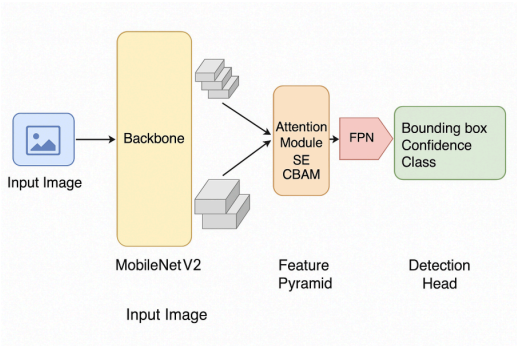


Figure 1. Structure diagram of this method

3.1. Action chunking transformer

The proposed model adopts modular design, including input layer, lightweight backbone network, attention enhancement module, feature fusion structure and output detection head. After image input, multi-level features are first extracted through backbone network, then the attention mechanism is applied to strengthen the response of crucial areas, and then multi-scale feature combination is used to integrate information at various levels. Finally, the detection head outputs the category, location and confidence of the target.

3.2. Lightweight backbone network

To efficiently minimize the model's computation load and storage usage during the inference stage, we selected MobileNetV2 as the backbone feature extraction network in the design. MobileNetV2 introduces two key technologies, namely, inverted residual structure and deep separable convolution. This structure could greatly reduce the number of model parameters and calculative complexity when maintaining a high feature expression ability. The network structure is especially well-suited for environments with limited resources., such as embedded devices and mobile terminals. In practical applications, MobileNetV2 can extract multi-level feature maps. This paper selects three representative scales for subsequent target detection operations, and perceives and locates targets

DOI: 10.54254/2755-2721/2025.AST26503

with three different sizes: small, medium, and large. This multi-scale feature extraction method helps to improve the generalization ability and detection accuracy of the model in complex scenes. However, due to the lightweight nature of MobileNetV2's own structure, its ability to extract fine-grained information from high-resolution images is limited, especially when dealing with images with complex textures or small targets, it is prone to the problem of insufficient feature representation. Therefore, in order to further improve the detail perception and spatial expression ability of the model, we introduced the attention mechanism and multi-scale feature fusion module in the subsequent network to enhance and optimize the backbone extraction results, so as to make up for its disadvantages in detail extraction.

3.3. Attention mechanism module

In order to enhance the model's focus on key target areas, this paper accesses the channel and spatial attention module after the feature map output from the backbone network, including SE module (Squeeze and Exception) and CBAM (Convolutional Block Attention Module). SE module can suppress redundant information through channel weighting, while CBAM further combines spatial attention, and has stronger responsiveness to small targets and low contrast areas. This module has minimal computational overhead, is suitable for lightweight model embedding, and can enhance detection accuracy while keeping the latency almost unchanged..

3.4. Multi-scale feature fusion

In order to adapt to the scale difference of the target in the actual scene, this paper uses a simplified feature pyramid structure (FPN) for feature fusion. The specific method is: the high-level feature map is upsampled, spliced with the low-level feature map, and then fused through convolution To generate a feature map that integrates semantic information from multiple levels. Different from the complex path fusion of traditional PANet or YOLOv5, this paper only adopts two-level fusion strategy to reduce model complexity and improve reasoning efficiency. It reduces network redundancy while maintaining sufficient context information. Although the simplified structure sacrifices part of the context expression ability, it still performs stably when combined with the attention mechanism in small target detection tasks.

3.5. Detection head design

The detection head refers to YOLO series design, and performs regression prediction on each grid cell of the feature map. The output includes the position of the target frame (center point, width and height), the confidence level of the existence of the target and its category. The output tensor dimension is $S \times S \times (5+C)$, where S is the size of the feature map, 5 is the border parameter and confidence, and C is the number of categories. In terms of loss function, the weighted combination of location regression (MSE), target confidence (binary cross entropy) and category classification (cross entropy) is adopted. The overall design can adapt to different platforms and resolution requirements, with good deployment flexibility.

DOI: 10.54254/2755-2721/2025.AST26503

4. Experiment

4.1. Experimental setup

To assess the performance of our efficient target detection model (HiT-Det) based on lightweight network and attention mechanism, we designed two typical target detection tasks: (1) Container Detection, and (2) Highway Detection. These two tasks represent actual scenarios with different complexity, and are used to test the generalization ability and robustness of the model in different scenarios. In the experiment, we used standard public data sets (such as Pascal VOC and COCO) to ensure that the task environment can provide accurate status feedback. At the same time, we introduced some uncertainties in the real scene, such as different lighting conditions, partial occlusion and background interference, to test the stability of the model in the changeable environment. The data set of each task adopts a unified standard preprocessing method in the training and testing stages to ensure the consistency of experimental results. In addition, we imposed a small disturbance on the target position in the experiment to further evaluate the adaptability of the model to environmental changes.

4.1.1. Container detection

This task is designed to assess the model's ability to detect container targets in complex industrial environments. As a large structural object, a container usually has a regular geometric appearance, but its color and texture are highly similar to the surrounding environment (such as storage yard, wharf facilities or transport vehicles), which significantly increases the difficulty of separability between the target and the background. In this experiment, the model needs to simultaneously detect multiple target areas related to containers from the image, including the main structure, number (ID) identification, enterprise logo, etc. To enhance detection performance, we incorporated a lightweight backbone network and an attention mechanism, enabling the model to concentrate on salient feature areas, so as to more accurately distinguish diversity boxing from complex backgrounds. In terms of evaluation, we calculated the detection accuracy of each type of target, and took mAP (mean Average Precision) as the overall performance index to comprehensively evaluate the overall performance of the model with respect to both accuracy and recall. In addition, we compared the performance of the model under different lighting conditions to verify its generalization ability and stability in complex scenes.

4.1.2. Highway detection

This task simulates a typical highway traffic scene, which contains multiple traffic targets in different motion states, such as trucks, cars, motorcycles, etc. Because these targets are often densely arranged and fast moving, the real-time and accuracy of the target detection model are required. In this experiment, the model needs to quickly and accurately identify all kinds of vehicles under complex background and different perspectives, and output its accurate boundary box position and category label for each target. In addition, the size distribution of traffic objects in the expressway scene is significantly different, and the size difference between large vehicles (such as trucks) and small objects (such as motorcycles) in the image is large, especially for the detection of small objects, with more challenges. Therefore, this experimental task is not only used to verify the detection speed and stability of the model, but also to evaluate its performance in small target detection and its adaptability to multi-scale targets. Through this task, we can more comprehensively

DOI: 10.54254/2755-2721/2025.AST26503

analyze the practicability and robustness of the model in a highly dynamic, target intensive and scale diversified environment.

4.2. Comparison method

To thoroughly assess the effectiveness of the HiT Det model, we selected five existing mainstream target detection methods as a comparison, and systematically sorted out the characteristics of each model, including the classic YOLOv5s, SSD300, and lightweight MobileNetV2 and Efficient Net lite. These methods represent the traditional convolutional neural network basic detection methods and the lightweight architecture emerging in recent years.

First, YOLOv5s, as a lighter version of YOLO series, uses CSPMarknet53 as the backbone network and combines feature fusion modules such as FPN and PANet, which is suitable for real-time target detection. However, due to the limited network size, it is relatively inaccurate in complex scenes and small target detection. In contrast, RetinaNet introduces the FPN structure for multi-scale feature fusion, and uses Focal Loss to address problem of category imbalance, which improves the detection performance of small targets. However, this method is a little slower in reasoning speed and is not suitable for strict real-time applications. The next is SSD300, which uses a single scale feature map for target detection, combined with the multi-scale default anchor frame detection scheme, suitable for processing the detection of large targets. Next is MobileNetV2, which builds a network structure with deeply separable convolutions, greatly decreasing the number of parameters and making it well-suited for resource-limited edge devices. Although the model is small, its feature extraction ability is relatively weak, which affects the overall detection accuracy, especially in complex scenes. EfficientNet lite uses a compound scaling strategy to balance the size and accuracy of the model. It performs well on medium scale datasets, but it has the problem of missing detection in target intensive or dynamic background environments.

Table 1. Comparison between HiT-Det and other algorithms

Algorithom	Container Detection	Highway Detection
YOLOv5s	72%	69%
SSD300	65%	60%
MobileNetV2	58%	54%
EfficientNet-lite	75%	73%
RetinaNet	78%	75%
HiT-Det (Ours)	85%	81%

5. Conclusion

This paper proposes an efficient target detection method based on lightweight network and attention mechanism, which aims to improve the speed, accuracy and deployability of target detection, especially in the resource constrained environment. By using a lightweight MobileNetV2 backbone network, introducing deep separable convolution (DWConv) to reduce the computational burden, and combining attention modules (such as SE) to dynamically weight features, this method significantly reduces the computational complexity and parameters of the model while maintaining high accuracy. In the experiment part, we verified the advantages of this method in multiple performance indicators by comparing classical target detection models (such as YOLOv5s and SSD300). Specifically, this method has successfully reduced the size of the model and improved the

DOI: 10.54254/2755-2721/2025.AST26503

reasoning speed while ensuring the detection accuracy equivalent to the baseline model. Particularly , the detection ability of small targets has been significantly improved. Experimental results show that the introduction of feature fusion and attention mechanism can greatly enhance the robustness and detection performance of the model against complex backgrounds and targets of different sizes. In addition, ablation experiments also verified the effectiveness of each innovative module, further proving the key role of lightweight design and attention mechanism in enhancing the performance of target detection model. The detection framework proposed in this paper has high flexibility and scalability. It can freely replace or combine lightweight backbone and attention modules according to different needs, and adapt to a variety of practical application scenarios. Overall, this study offers an efficient, adaptable, and highly practical approach for target detection tasks. It can achieve rapid and accurate target detection in the environment with limited resources, and holds wide-ranging application potential, especially in the fields of autopilot, unmanned aircraft, monitoring, and so on. Future work can further optimize the network structure, explore more lightweight strategies and different types of attention mechanisms, and improve the detection ability and generalization of the model.

References

[1] Szeliski R. Computer vision: algorithms and applications [M]. Springer Nature, 2022.
[2] Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions [J]. Journal of big Data, 2021, 8(1): 53.
[3] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas [J]. Machine learning and knowledge extraction, 2023, 5(4): 1680-1716.
[4] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector [C]//European conference on computer vision. Cham: Springer International Publishing, 2016: 21-37.
[5] Jiang L, Nie W, Zhu J, et al. Lightweight object detection network model suitable for indoor mobile robots [J]. Journal of Mechanical Science and Technology, 2022, 36(2): 907-920.
[6] Chen Y, Dai X, Chen D, et al. Mobile-former: Bridging mobilenet and transformer [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5270-5279.
[7] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856. Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies [J]. Journal of Machine Learning Research, 2016, 17(39): 1-40.
[8] Han Q, Fan Z, Dai Q, et al. On the connection between local attention and dynamic depth-wise convolution [J]. arXiv preprint arXiv: 2106.04263, 2021. 2Levine S , Pastor P , Krizhevsky A , et al.Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection [J].International Journal of Robotics Research, 2016, 37(4-5): 421-436.DOI: 10.1007/978-3-319-50115-4_16.
[9] Ab Wahab M N, Nazir A, Ren A T Z, et al. Efficientnet-lite and hybrid CNN-KNN implementation for facial expression recognition on raspberry pi [J]. IEEE Access, 2021, 9: 134065-134080.
[10] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[11] Ghiasi G, Lin T Y, Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7036-7045.
[12] Xiao L, Li W, Yao S, et al. High-precision and lightweight small-target detection algorithm for low-cost edge intelligence [J]. Scientific Reports, 2024, 14(1): 23542.
[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[14] Liu Y, Cao S, Lasang P, et al. Modular lightweight network for road object detection using a feature fusion approach [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(8): 4716-4728.
[15] Liu Y, Cao S, Lasang P, et al. Modular lightweight network for road object detection using a feature fusion approach [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(8): 4716-4728.