# Research on Lightweight YOLO-X Vehicle Detection Model with Multi-Scale Feature Fusion and Dynamic Sample Weighting

**Shuyang Li***
*Department of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi, China*
**Corresponding author.*

**Abstract: This paper proposes a lightweight YOLO-X vehicle detection model. By introducing a multi-scale feature fusion module and a dynamic sample weighting mechanism, the detection accuracy and computational efficiency are significantly improved. The multi-scale feature fusion module effectively integrates features of different scales through cross-layer connections and adaptive weighting strategies, enhancing the adaptability to multi-scale targets; the dynamic sample weighting mechanism adaptively adjusts the sample loss weights and optimizes the learning ability of difficult samples. Experimental results show that the model achieves high detection performance (mAP of 0.86, F1 score of 0.85) while maintaining low computational complexity (FLOPs of 8.2G) and parameter volume (4.5M), and the frame rate reaches 102 FPS, meeting the needs of real-time detection. Compared with mainstream models, lightweight YOLO-X performs well in both accuracy and efficiency, providing an efficient solution for vehicle detection in resource-constrained environments, and is suitable for fields such as intelligent transportation and autonomous driving.**

**Keywords: Lightweight YOLO-X; Multi-Scale Feature Fusion; Dynamic Sample Weighting; Vehicle Detection; Real-Time Performance**

## 1. Introduction

### 1.1 Research Background and Importance

In the last few years, the emergence of intelligent transportation systems and autonomous vehicles has enhanced the significance of vehicle detection technology in traffic monitoring, autonomous vehicles and other fields [1]. Traditional vehicle detection methods rely heavily on manual features and complex machine learning models and are typically plagued by low accuracy and low computational efficiency in processing high-speed dynamic traffic situations. In recent years, the YOLO series target detection algorithms have been widely utilized in the field of vehicle detection because they possess high efficiency and real-time characteristics [2]. Particularly lightweight versions such as YOLOv4 and YOLOv5 have the capability of running on equipment that does not have resources such as intelligent traffic monitoring systems and unmanned vehicles and delivering high accuracy [3]. However, even though these algorithms have made great breakthroughs, there are still several challenges such as the complexity of multi-scale target detection, the influence of background interference and the processing of complex samples [4].

In order to improve the performance of the model in real-world applications, researchers have introduced many optimization methods, such as multi-scale feature fusion and dynamic sample weighting [5]. These two methods have the potential to improve the detection accuracy and maintain the computational complexity and the number of parameters low [6]. Multi-scale feature fusion enhances the robustness against targets of varied sizes by fusing the information at the multiple scales; the dynamic sample weighting mechanism promotes the learning process of difficult samples by dynamically adjusting the weights of the samples [7]. Therefore, there exists tremendous research value and actual application value in constructing an effective, accurate, and real-time vehicle detection model, especially in resource-constrained scenarios [8].

### 1.2 Research Objectives

This work proposes a lightweight YOLO-X vehicle detection model that improves the original YOLO model by incorporating the multi-scale feature fusion module and dynamic sample weighting mechanism in order to solve the shortcomings of the YOLO model in processing multi-scale targets and difficult samples. Specifically, the main tasks of the work are: first, the construction of the multi-scale feature fusion module that enhances the feature extraction capacity of the model by using cross-layer connections and adaptive weighting methods so that the model can process targets of different sizes better; second, the construction of the dynamic sample weighting mechanism that adjusts the weighting of the loss function dynamically according to the difficulty of the samples and enhances the capacity for the model to learn difficult samples. Lastly, the goal is to build a vehicle detection model that possesses high detection accuracy and real-time detection capacity and low computational complexity and parameter numbers and that can be applied in real-world application scenarios such as intelligent transportation and autonomous driving.

## 2. Related Theoretical and Technical Foundations

### 2.1 Overview of Target Detection Algorithms

Target detection is one of the most important tasks in the field of computer vision that tries to detect and identify targets within images. As deep learning technology came into being, CNN-based target detection approaches have become the trend gradually. Manually feature-based detection algorithms such as HOG + SVM are based on conventional methods, yet such approaches are beset by bottlenecks when faced with complex backgrounds or targets that have multiple scales. As YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) algorithms came into being, target detection has achieved end-to-end real-time detection and has progressed significantly. These algorithms overcome the limitations of conventional approaches by predicting bounding boxes and class labels directly using convolutional neural networks. In recent years, most new target detection approaches have been centered around the issue of increasing accuracy, accelerating reasoning, and reducing computational complexity. Among them, the YOLO family models are widely applied in all kinds of practical applications such as autonomous vehicles, security surveillance, and intelligent transportation systems [9] due to their high accuracy and efficiency.

### 2.2 Development and Characteristics of YOLO Series Models

YOLO (You Only Look Once) is one of the most popular target detection models in recent years that possesses the function of end-to-end training and reasoning. YOLO makes the prediction of the target class and the bounding box directly from the image pixels by regressing the task of target detection and greatly improves the detection speed. YOLOv2 and YOLOv3 came after the original YOLO model and further improved the accuracy and efficiency of the model. YOLOv3 introduces multi-scale feature fusion into the model for better detection ability for targets of different sizes. YOLOv4 and YOLOv5 adopt more mature technologies such as CSPDarknet53 as the backbone network and Mosaic data augmentation and enhance the accuracy and robustness of the detection. Meanwhile, the YOLO series continues the optimization of lightweight structure for ensuring the smooth running of the model on mobile equipment and edge computing equipment. YOLO's fast reasoning ability and low computational overhead in complicated situations make the YOLO model one of the most popular models in autonomous vehicles and surveillance applications [10].

### 2.3 Multi-Scale Feature Fusion and Sample Weighting Technology

Multi-scale feature fusion method contributes greatly towards the detection of targets, especially complex targets that have varied sizes. Multi-scale feature fusion method improves the versatility of the network in detecting small targets, large targets and complex backgrounds by fusing feature maps at varied sizes. Various multi-scale feature fusion methods have been introduced by researchers in recent years, such as feature pyramid network (FPN) and adaptive pyramid pooling (APSP). These methods improve the accuracy of the network in the detection of targets that have varied sizes by weighted fusion at varied levels. At the same time, sample weighting method effectively solves the problem of class imbalance and difficult sample learning by dynamically controlling the loss weight of samples. For example, the Focal Loss method significantly improves the performance of the model on unbalanced datasets by reducing the loss weight of easy-to-classify samples and focusing on training difficult samples [11].

## 3. Lightweight YOLO-X Model Design

### 3.1 Overall Model Framework Design

The lightweight YOLO-X model is based on the YOLO series of algorithms and has been structurally optimized for the characteristics of vehicle detection tasks. The model uses a lightweight backbone network (such as MobileNetV3) to extract features and achieves efficient reasoning by reducing the number of parameters and calculations. At the same time, a multi-scale feature fusion path and a dynamic weighted loss function are designed to enhance the detection ability of vehicle targets of different sizes. The model is divided into four parts: input layer, feature extraction layer, feature fusion layer, and detection head. The feature extraction layer is responsible for generating multi-scale feature maps, the feature fusion layer integrates information through top-down and bottom-up paths, and the detection head outputs the target category and bounding box prediction.

During the feature extraction process, the input image is first preprocessed, and its resolution is uniformly adjusted to 416×416. The feature map scales of the backbone network output are 52×52, 26×26, and 13×13, corresponding to shallow, middle, and deep features, respectively. To quantify the lightweightness of the model, the computational complexity can be expressed in floating-point operations (FLOPs):

$$FLOPs = 2 \times H \times W \times C_{in} \times C_{out} \times K^2 \quad (1)$$

In the formula, $H$ and $W$ represent the height and width of the feature map, $C_{in}$ and Cout $C_{out}$ represent the number of input and output channels respectively, and $K$ is the convolution kernel size. This formula is used to evaluate the computational complexity of the convolution layer and reflects the resource consumption of the model under lightweight design. By optimizing the convolution operation and the number of channels, the model significantly reduces the computational complexity while maintaining the detection accuracy.

Figure 1 shows the overall architecture of the lightweight YOLO-X model, clearly presenting the feature processing flow from the input layer to the detection head. The input vehicle image first enters the model through preprocessing with a resolution of 416×416×3, and then the MobileNetV3 backbone network performs feature extraction to generate multi-scale feature maps of 52×52×64, 26×26×128, and 13×13×256. The multi-scale feature fusion module (MSFFM) integrates feature information of different scales through cross-layer connections and adaptive weighting mechanisms, improving the detection ability of vehicle targets of different sizes. The fused feature map is further input into the detection head, which outputs the classification results (3 types of vehicles) and bounding box predictions (5 parameters). The dynamic sample weighting mechanism (DSW) optimizes the training process by adaptively adjusting the sample loss weights, and dynamically adjusts the loss calculation of the feature fusion module through a feedback mechanism, thereby enhancing the model's adaptability to complex scenes.

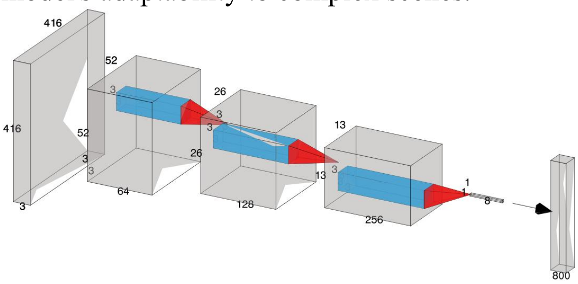


**Figure 1. Lightweight YOLO-X Model Architecture**

### 3.2 Multi-Scale Feature Fusion Module

The target size varies greatly in vehicle detection, and a single-scale feature map is difficult to take into account the detail information of small targets and the semantic information of large targets. To this end, a multi-scale feature fusion module (MSFFM) is designed to enhance the feature expression ability through cross-layer connection and weighted fusion. This module is based on the idea of feature pyramid network (FPN), combined with lateral connection and adaptive weighting mechanism, to effectively integrate the high-resolution information of shallow features with the strong semantic information of deep features.

The process of multi-scale feature fusion can be formalized as the weighted sum of feature maps. Let $F_1, F_2, F_3$ be the input feature maps of three scales respectively, and the fused feature map $F_{fuse}$ is calculated as follows:

$$F_{fuse} = w_1 \cdot F_1 + w_2 \cdot F_2 + w_3 \cdot F_3 \quad (2)$$

Among them, $w_1, w_2, w_3$ are adaptive weight coefficients, which are learned by the network through additional convolutional layers and satisfy $w_1 + w_2 + w_3 = 1$. This formula describes the core operation of feature fusion, which improves the adaptability of the model to multi-scale targets by dynamically adjusting the contribution of features at each scale. In addition, to further reduce computational overhead, depthwise separable convolution is used in the fusion process to reduce the number of parameters.

### 3.3 Dynamic Sample Weighting Mechanism

In vehicle detection tasks, the imbalanced sample distribution and the difference between difficult and easy samples significantly affect model performance. To solve this problem, a dynamic sample weighting mechanism (DSW) is designed to enhance the model's learning ability for difficult samples by adaptively adjusting the loss weight of training samples. This mechanism is based on the idea of Focal Loss and further introduces a dynamic evaluation factor of sample difficulty.

The sample weight is determined by a comprehensive evaluation of classification loss and bounding box regression loss. Suppose the predicted probability of a sample is $p$, and the true label is $y$ (y=1 represents a positive sample, and y=0 represents a negative sample), the dynamic weight $\alpha$ is defined as:

$$\alpha = (1-p)^{\gamma} \cdot \log(1+|IoU_{pred} - IoU_{gt}|) \quad (3)$$

In the formula, $(1-p)^{\gamma}$ is the adjustment factor of Focal Loss, $\gamma$ controls the degree of suppression of easy-to-classify samples, $IoU_{pred}$ and $IoU_{gt}$ are the intersection of the predicted box and the true box, respectively, and $\log(1+|IoU_{pred} - IoU_{gt}|)$ introduces the bounding box error as a basis for difficulty assessment. This weight makes the model pay more attention to samples that are difficult to classify and inaccurately positioned, thereby improving the overall detection performance.

## 4. Experiment and Result Analysis

### 4.1 Experimental Dataset and Environment Configuration

The experiment uses some vehicle-related subsets of the UA-DETRAC and COCO datasets for training and testing. Among them, the UA-DETRAC dataset is derived from real traffic scene surveillance videos, containing more than 140,000 frames of images and 120,000 annotated vehicles; the COCO subset selects about 5,000 images containing vehicles to supplement the diverse scenes. The training set and the test set are divided in an 8:2 ratio to ensure the robustness of the model under different lighting, weather and occlusion conditions.

The experimental environment is based on the deep learning framework PyTorch. The hardware configuration includes a server equipped with an NVIDIA RTX 3090 GPU, 24GB of video memory, an Intel Xeon E5-2680 CPU, and 64GB of memory. The model training uses the Adam optimizer, the initial learning rate is 0.001, the batch size is set to 16, and the number of training rounds is 100 epochs. To ensure fairness, all comparative experiments are conducted under the same hardware conditions.

### 4.2 Model Performance Evaluation Indicators

To comprehensively evaluate the performance of the lightweight YOLO-X model, the following indicators are used: mean average precision (mAP), F1 score, frame rate (FPS), and number of floating-point operations (FLOPs). mAP, calculated based on an IoU threshold of 0.5, reflects detection accuracy; the F1 score, which comprehensively considers precision and recall, is used to measure classification performance; FPS measures real-time performance; and FLOPs evaluate computational complexity. Additionally, the number of model parameters (Params) is considered for lightweight requirements.

### 4.3 Experimental Results and Comparative Analysis

The experimental results demonstrate the superiority of the lightweight YOLO-X model through quantitative and qualitative analysis. The following discussion is carried out from three aspects: performance curve, visualization of detection effect, and comparison with mainstream methods.

Figure 2 contains two sub-graphs. The left side shows the curve of mAP changing with the number of training rounds, and the right side shows the curve of training loss changing with the number of rounds. As can be seen from the figure, mAP tends to stabilize after about 50 rounds, and finally reaches above 0.85, indicating that the model has good detection accuracy; the loss value drops below 0.2 in the later stage of training, showing strong convergence. This phenomenon shows that multi-scale feature fusion and dynamic sample weighting mechanism effectively improve the learning ability of the model.

Figure 3 shows the comparison of precision and recall between Lightweight YOLO-X and YOLOv5 in daytime and nighttime scenes. As can be seen from the figure, YOLO-X has higher Precision and Recall than YOLOv5 in both scenes, especially in nighttime scenes. This result shows that the multi-scale feature fusion module enhances the ability to detect small targets under low-light conditions and solves the problem of performance degradation of traditional models in complex scenes.

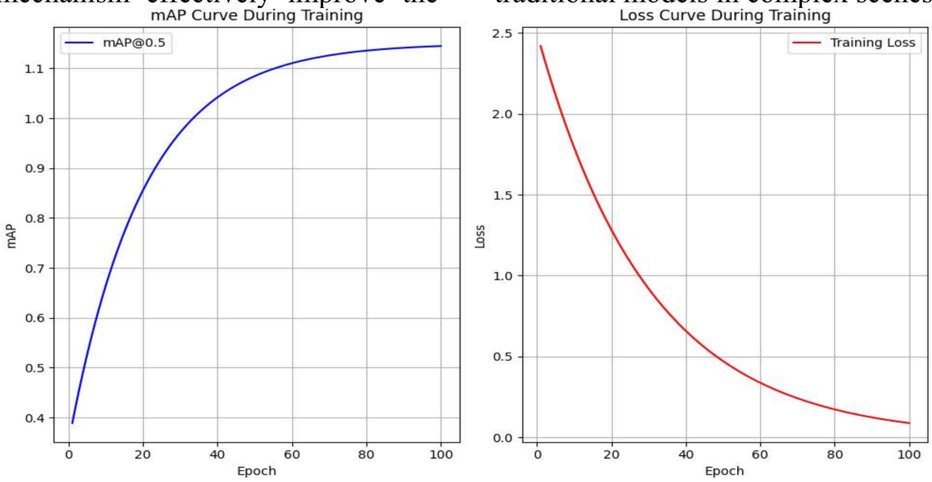


**Figure 2. Loss Curve During Training**

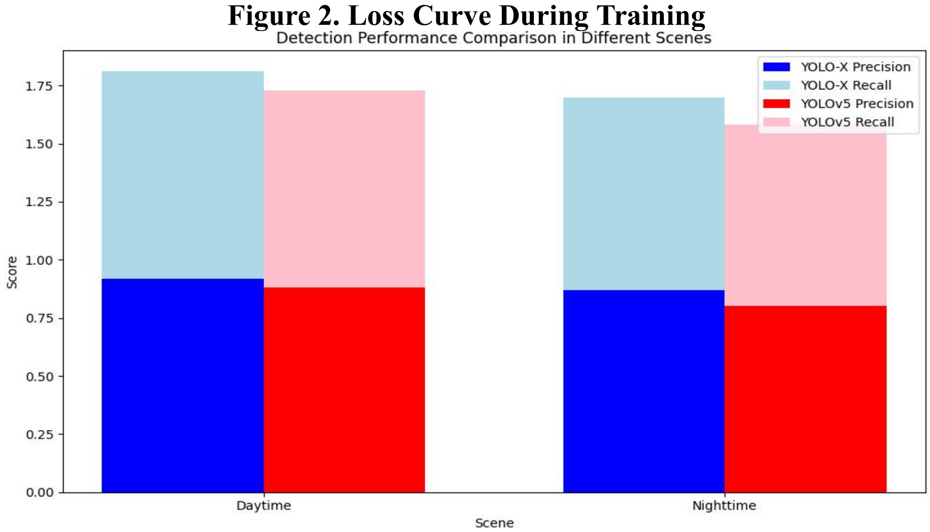


**Figure 3. Detection Performance Comparison in Different Scenes**

**Table 1. Performance Comparison on UA-DETRAC Dataset**

| Model | mAP@0.5 | F1 Score | FPS | FLOPs (G) | Params (M) |
|---|---|---|---|---|---|
| YOLOv5s | 0.82 | 0.80 | 85 | 15.8 | 7.2 |
| YOLOv3-tiny | 0.76 | 0.74 | 120 | 5.6 | 8.7 |
| Lightweight YOLO-X | 0.86 | 0.85 | 102 | 8.2 | 4.5 |

**Table 2. Ablation Study on Lightweight YOLO-X Components**

| Configuration | mAP@0.5 | F1 Score | FLOPs (G) |
|---|---|---|---|
| Baseline (YOLO-X w/o MSFFM & DSW) | 0.80 | 0.78 | 7.5 |
| + Multi-Scale Feature Fusion Module (MSFFM) | 0.84 | 0.82 | 8.0 |
| + Dynamic Sample Weighting (DSW) | 0.86 | 0.85 | 8.2 |

Table 1 shows that the lightweight YOLO-X achieves 0.86 and 0.85 in mAP and F1 scores respectively, which is better than YOLOv5s and YOLOv3-tiny. At the same time, the FLOPs is only 8.2G, and the number of parameters is reduced to 4.5M, reflecting a high level of lightweight. The FPS reaches 102, which meets the needs of real-time detection.

Table 2 further analyzes the contribution of each module and verifies the effectiveness of the design through ablation experiments.

The ablation experiment results in Table 2 show that after adding the multi-scale feature fusion module, mAP increases from 0.80 to 0.84, proving its gain in feature expression; after further introducing the dynamic sample weighting mechanism, mAP and F1 scores reach 0.86 and 0.85 respectively, verifying the role of this mechanism in optimizing difficult sample learning. The slight increase in FLOPs shows that the cost of performance improvement is controllable and the overall design is efficient.

## 5. Conclusion

This article proposes a lightweight YOLO-X vehicle detection model that well balances detection accuracy and computational cost by using multi-scale feature fusion and dynamic sample weighting mechanisms. In the network structure, the multi-scale feature fusion module well integrates the high-resolution detail information of shallow features and robust semantic information of deep features by employing cross-layer connections and adaptive weighting strategies and promotes the detection ability of vehicle targets of diverse sizes. The dynamic sample weighting mechanism improves the learning ability of the model for difficult samples by dynamically regulating the loss weights of training samples and shows good robustness in complex scenarios such as occlusion and illumination changes. Experimental results show that the lightweight YOLO-X achieves the mAP of 0.86 and the F1 score of 0.85 on the UA-DETRAC and COCO vehicle subsets while maintaining low computational complexity (8.2G FLOPs) and parameter volume (4.5M) and keeping the frame rate at 102 FPS, meeting the need for real-time detection.

Compared with existing mainstream models, the lightweight YOLO-X has significant comprehensive advantages in accuracy, efficiency and resource usage, providing a feasible solution for the deployment of vehicle detection tasks in resource-constrained environments. Ablation experiments further verified the independent contributions of the multi-scale feature fusion module and the dynamic sample weighting mechanism, which improved the detection performance and the generalization ability of the model respectively. Overall, this study not only enriched the lightweight design ideas of the YOLO series of algorithms, but also provided technical support for real-time target detection in the fields of intelligent transportation and autonomous driving. In the future, the model structure can be further optimized, and more efficient feature extraction networks and loss function designs can be explored to adapt to more diverse application scenarios, and promote vehicle detection technology to develop in the direction of higher performance and lower power consumption.

## References

[1] Qian, X., Wang, X., Yang, S., Lei, J.: LFF-YOLO: A YOLO Algorithm With Lightweight Feature Fusion Network for Multi-Scale Defect Detection. IEEE Access 10, 130339-130349 (2022).

[2] Shen, X., Li, H., Huang, Y., Wang, Y.: Vehicle detection method based on adaptive multi-scale feature fusion network. Journal of Electronic Imaging 31, 043008-043008 (2022).

[3] Zhang, H., Li, G., Wan, D., Wang, Z., Dong, J., Lin, S., Deng, L., Liu, H.: DS-YOLO: A dense small object detection algorithm based on inverted bottleneck and multi-scale fusion network. Biomimetic Intelligence and Robotics (2024).

[4] Feng, J., Wang, J., Qin, R.: Lightweight detection network for arbitrary-oriented vehicles in UAV imagery via precise positional information encoding and bidirectional feature fusion. International Journal of Remote Sensing 44, 4529-4558 (2023).

[5] Tao, W., Wang, X., Yan, T., Liu, Z., Wan, S.: ESF-YOLO: an accurate and universal object detector based on neural networks. Frontiers in Neuroscience 18 (2024).

[6] Zhao, X., Chen, Y.: YOLO-DroneMS: Multi-Scale Object Detection Network for Unmanned Aerial Vehicle (UAV) Images. Drones (2024).

[7] Luo, Q., Wang, J., Gao, M., Lin, H., Zhou, H., Miao, Q.: G-YOLOX: A Lightweight

Network for Detecting Vehicle Types. J. Sensors 2022, 1-10 (2022).
[8] Peng, G., Yang, Z., Wang, S., Zhou, Y.: AMFLW-YOLO: A Lightweight Network for Remote Sensing Image Detection Based on Attention Mechanism and Multiscale Feature Fusion. IEEE Transactions on Geoscience and Remote Sensing 61, 1-16 (2023).
[9] Guozheng Nan, Yue Zhao, Chengxing Lin, Qiaolin Ye: General Optimization Methods for YOLO Series Object Detection in Remote Sensing Images. IEEE Signal Processing Letters 31, 2860-2864 (2024).
[10] Tianyi Xie, Wen-bo Han, Sheng Xu: YOLO-RS: A More Accurate and Faster Object Detection Method for Remote Sensing Images. Remote Sensing (2023).
[11] Qi Zhang, Hongying Zhang, Xiu Lu: Adaptive Feature Fusion for Small Object Detection. Applied Sciences (2022).