

Road Traffic Object Detection System for Severe Weather Scenarios Based on YOLOX

Hao Jiang

*College of Computer Science, Chengdu University, Chengdu, China
202310112113@cdu.edu.cn*

Abstract. Road traffic object detection is the core perceptual task in the current field of intelligent transportation. The You Look Only Once (YOLO) series algorithms are widely used for real-time detection due to their end-to-end inference, fast speed, high accuracy, and ease of deployment. However, existing public datasets are mostly collected in sunny weather, and harsh conditions such as rainy, foggy, snowy, and nighttime can lead to image quality degradation, blurred object edges, reduced contrast, and insufficient lighting, resulting in a significant decrease in detection performance. In response to the above issue, this paper uses YOLOX as the basic model, introduces SE module and small object detection head method to improve the detection ability of the model in complex environments, and verifies the effectiveness of the proposed method on the Adverse Conditions Dataset with Correspondences (ACDC) dataset. The above improvements resulted in a mAP value of only 21.86% on the ACDC dataset, a decrease of about 0.41% compared to the baseline method, but an increase of 1.8% in small object AP values and 2% in car class detection. In the future, further optimization can be attempted from the perspective of changing the introduction method.

Keywords: Object detection, YOLOX, Adverse weather, Feature enhancement, Data augmentation

1. Introduction

In recent years, with the rapid development of fields such as autonomous driving, intelligent transportation, and smart cities, road traffic object detection technology has become one of the important directions of computer vision research. Object detection algorithms need to be able to accurately identify targets such as vehicles, pedestrians, traffic signs, and cyclists on the road, providing reliable information for subsequent target tracking, behavior prediction, and path planning.

At present, significant achievements have been made in deep learning based object detection algorithms, among which the You Look Only Once (YOLO) series algorithms are widely used in real-time object detection tasks due to their advantages of end-to-end inference, fast detection speed, high accuracy, and easy deployment. YOLOX, as an academic modified version of YOLO algorithm released in 2021, is based on YOLOv3 by Megvii, a Chinese company. It integrates the Anchor Free detection mechanism and decoupling detection head, which were only proposed but not applied at

that time, to further improve detection accuracy while maintaining high detection speed [1]. This project has 2.5k forks and 10.5k stars on GitHub, and is extremely popular in the community [2]. Therefore, this method has become the fundamental detection model for this project.

However, when the environment experiences severe weather conditions such as rainy, foggy, snowy, and nighttime, the image quality significantly decreases, and problems such as blurred target edges, reduced contrast, and insufficient lighting can lead to a significant decline in detection performance. In their review study, Shen Ruifan et al. found that at the current level of industry research, there are relatively few image processing algorithms and detection improvement methods for various meteorological conditions. However, there are many effective image processing algorithms for single meteorological conditions, especially for foggy weather [3]. Wang et al. proposed an HCD dehazing method in their research, which showed significant improvements in PSNR, SSIM, and other indicators [4]. The dehazing method demonstrated the widespread research of image processing algorithms in the current industry under single weather conditions. At the same time, there have not been many projects that have used YOLOX as a baseline method for object detection in adverse weather conditions. Zhang Haining used YOLOv5 as the basic framework in his research, and used its newly proposed backbone network structure and modules to perform targeted upgrades for various severe weather recognition [5]. Similarly, Aloufi et al. used YOLOv5 as the foundational framework in their research and proposed a novel multi-objective model that integrates image quality assessment SRGAN, by using the DAWN dataset, achieve an average accuracy of 74.6% [6]. Both studies mentioned above used YOLOv5 for in-depth research. Although Luo Kehao et al. proposed a detection algorithm based on YOLOX for complex weather, which added rain, snow, and motion blur conditions to the TT100K, CCTSDB, and GTSDB datasets, the improved YOLOX-s algorithm improved mAP accuracy by 2.1%, 0.1%, and 5.7%, respectively. But this algorithm is only applicable to traffic signs [7]. Therefore, studying the detection capability of YOLOX under adverse weather conditions has important research value and practical application significance.

This project focuses on the detection of road traffic targets in adverse weather conditions, using YOLOX as the basic model. By introducing feature enhancement and data augmentation strategies, the detection performance of the model in complex environments is improved, and the effectiveness of the algorithm is verified using the Adverse Conditions Dataset with Correspondences (ACDC) dataset.

2. Research methods

2.1. Dataset introduction

Firstly, organize the severe weather dataset and construct training, validation, and testing sets based on official divisions. The training set is used for model parameter learning, the validation set is used for hyperparameter adjustment, and the test set is mainly used for final performance evaluation of the model. After conducting research, the paper has decided to use the ACDC dataset for training.

ACDC It is a publicly available dataset for the perception of harsh visual scenes in autonomous driving, launched by the Swiss Federal Institute of Technology Zurich at ICCV2021. The dataset contains 4006 real road images, uniformly covering four typical adverse visual conditions: fog, night, rain, and snow, and providing GPS matched clear daytime reference images of the same scene for each adverse scene image. The dataset provides 19 pixel level panoramic segmentation annotations compatible with Cityscapes, along with uncertainty mask labels for difficult to identify areas, supporting semantic segmentation, object detection, panoramic segmentation, and uncertainty

aware segmentation tasks. ACDC is often used as a standard evaluation benchmark for robust perception, image restoration, and cross domain transfer algorithms in adverse weather scenarios [8].

2.2. Analysis of adverse weather scenarios

2.2.1. Fog scene

Foggy weather can lead to increased light scattering in the air, decreased overall image contrast, blurred edges of distant targets, and greater difficulty in identifying small targets. As shown in Figure 1, foggy weather can significantly reduce the characteristics of long-distance cars, and the overall color of the photo becomes darker, causing interference with foreground information such as reduced background contrast.

For YOLOX, due to the weakening of target texture information, the feature extraction stage is prone to losing effective target information, thereby reducing detection accuracy.



Figure 1. Foggy image in the ACDC dataset [8]

2.2.2. Rainy scene

Raindrops, standing water, and rain curtains can produce a large amount of interference textures, affecting the integrity of the target contour. Meanwhile, road reflections can also cause false positives. As shown in Figure 2, the raindrops in the rainy day image cause morphological distortion of vehicles on the left side of the road, and there are reflections from the ground and objects inside the vehicle, which interfere with the ideal object detection and feature learning methods.

In addition, raindrops and wipers on the camera lens may cause local blurring, resulting in deviation in the positioning of the target bounding box.



Figure 2. Rainy images in the ACDC dataset [8]

2.2.3. Snowy scene

Snow covering the target, changes in background color, and snow on the ground can all reduce the discrimination between the target and the background. Especially between white vehicles and snowy backgrounds, there is a risk of unclear target edges. As shown in Figure 3, the contrast between the white box truck in front and the snow on the sky and ground is significantly lower, and the snow accumulation on the black vehicle on the right side has caused a significant change in the vehicle's exterior characteristics.



Figure 3. Snowy images in the ACDC dataset [8]

2.2.4. Night scene

Nighttime environments often suffer from issues such as insufficient lighting, uneven exposure, and strong light interference. For example, street lamps and car lights may produce glare, while distant targets may be completely hidden in dark areas. Nighttime object detection is usually one of the most difficult scenarios among the four types of weather. As shown in Figure 4, high-speed moving cars in nighttime images can cause ghosting due to glare, while foreground objects in dark areas blend into the dark background, resulting in a significant decrease in contrast.



Figure 4. Nighttime images in the ACDC dataset [8]

2.3. System boundary design

In order to ensure clear project objectives, the research scope of this project has been limited. In terms of detection principle, this system does not completely redesign the overall network structure. Instead, it fine tunes the network structure based on the existing mature YOLOX algorithm and conducts experimental analysis on other factors. This system focuses on two-dimensional image object detection, mainly focusing on road traffic targets, including typical traffic targets such as motor vehicles, non motor vehicles, pedestrians, etc. It does not involve other visual tasks such as lane detection, traffic signal recognition, and semantic segmentation. In terms of data sources, only the ACDC dataset was used for experiments, and data was not collected or annotated by oneself to ensure the reproducibility and comparability of experimental results. At the same time, this stage of work does not involve embedded hardware deployment, real-time optimization in vehicles, or large-scale engineering applications.

2.4. Evaluation metric

This article uses YOLOX to train an object detection model. The model takes road traffic scene images as input, extracts features through Backbone, performs multi-scale feature fusion through Neck, and finally completes object classification and bounding box prediction through Head.

After completing model training, use the validation set for inference testing, calculate evaluation metrics such as Precision, Recall, mAP, and analyze the impact of different weather conditions on detection performance. Precision measures the proportion of samples predicted as positive by the model that are truly positive. The calculation formula for this indicator is:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Among them, TP is the number of samples correctly predicted as positive cases, and FP is the number of negative samples incorrectly predicted as positive cases. The higher the precision, the less likely the model will detect foreground objects incorrectly.

Recall measures the proportion of all true positive cases successfully detected by the model. The calculation formula for this indicator is:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Among them, FN refers to the number of true positive cases that were not detected by the model, that is, missed targets. The higher the recall, the fewer labeled prospects the model misses.

AP is the area of the curve formed by the two major indicators of Precision Recall, used to evaluate the comprehensive detection performance of the model on a single category. MAP is the arithmetic mean of all categories of APs used to evaluate the overall detection performance of the model across all categories. The calculation formula for this indicator is:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (3)$$

3. Experimental design

3.1. Baseline method for direct training and testing

The overall structure of YOLOX is actually the same as most YOLO series, consisting of three parts: Backbone, Neck, and Head.

In YOLOX, Backbone is a pre trained convolutional neural network CSP-DarkNet-53 that acts as a feature extraction network to compute feature maps from input images. Neck connects Backbone and Head, consisting of a Feature Pyramid Network (FPN) and a Path Aggregation Network that combines low-level and high-level features. The Feature Pyramid Network generates feature maps and corresponding grids at multiple scales. Connect feature maps from the backbone layer and use them as inputs to the head at three different scales (1024, 512, and 256 channels). The decoupled detection head processes the aggregated features into three feature channels, including classification score (the category of each bounding box), regression score (the position and size of each bounding box), and IoU (the confidence score of each bounding box containing the object).

Figure 5 shows the basic YOLOX architecture, which for simplicity only displays a decoupling head at one of the three scales.

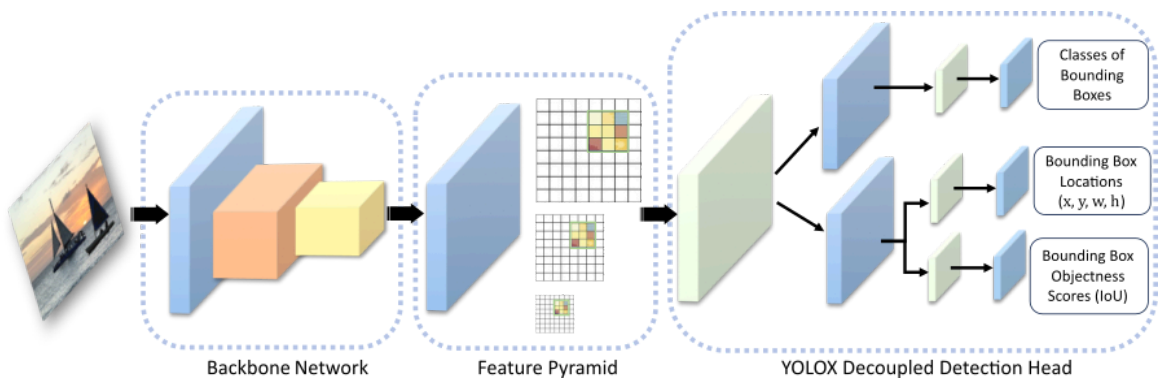


Figure 5. YOLOX basic network architecture [9]

The core innovation of YOLOX is the detection head and training method. The baseline method YOLOv3 of this method consists of a detection head responsible for classification, bounding box regression, and target confidence prediction simultaneously. And the algorithm splits it into

independent branches, allowing classification and localization to learn their own features separately, reducing task conflicts and improving detection accuracy. At the same time, abandoning the detection method that relies on preset anchor boxes to generate candidate boxes, using the position on the feature map as the prediction starting point, directly regressing the target center position and bounding box size, thereby reducing hyperparameters, simplifying the detection process, and improving the model's generalization ability.

Based on the above content, this article considers using the YOLOX-S model and its training weights, directly calling the script for training on this basis, training weights based on the Baseline method, and obtaining accuracy, recall mAP. Among them, the core hyperparameters for training are shown in the table 1:

Table 1. Core hyperparameters used in training

Name	Value	Name	Value
Depth coefficient	0.33(YOLOX-S)	Momentum	0.9
Width coefficient	0.50(YOLOX-S)	Mosaic probability	0.5
Training input size	640×640	MixUp probability	0.5
Verify input size	640×640	HSV enhancement probability	1.0
Maximum number of training epochs	100	Flip probability	0.5
Learning rate warm-up rounds	5	Rotation angle range	±10°
No enhanced rounds	30	Translation range	0.1
Verification interval	Perform verification every 10 rounds	Mosaic zoom range	(0.5, 1.5)
Basic learning rate	0.01/16.0 (per image)	Enable MixUp image feature enhancement	YES
weight decay	1e-4		

3.2. Lighting adaptive feature enhancement strategy

To solve the problem of feature expression degradation under complex lighting conditions such as strong backlighting and nighttime halos, I embedded SE channel attention modules after the last two cross stage local modules of Backbone based on the method proposed by Yang Rui et al.

The SE module effectively compresses spatial information through global average pooling, and then learns the importance weights of each channel through a two-layer fully connected network. Finally, it performs adaptive recalibration of the channel dimension on the feature map.

The channel attention weight calculation formula for the SE module is:

$$s = \sigma [W_2 (\delta (W_1 z))] \quad (4)$$

In the formula: $z \in \mathbb{R}^C$ is the channel description vector after global average pooling, where C is the number of channels (dimensionless, taken as 512); $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are respectively the weight matrices of two fully connected layers, where r is the compression ratio (taken as 16); $\delta (\cdot)$ is the ReLU activation function; $\sigma (\cdot)$ is the Sigmoid function that maps the

output value to $[0,1]$; $s \in \mathbb{R}^C$ is the final channel attention weight vector used to scale each channel of the input feature map element by element [10].

3.3. Small object detection head enhancement module

The YOLOX original structure consists of three detection scales, namely 80×80 , 40×40 , and 20×20 . The minimum receptive field corresponds to a feature map resolution of 80×80 . When the resolution of the target in the original image is less than 32×32 , a large number of effective features are lost after multiple downsampling, resulting in missed detection of small targets. I introduced a fourth detection branch at the Neck output based on the method mentioned in Yang Rui et al.'s article, using a 160×160 resolution feature map.

This branch is obtained by bilinear upsampling the P3 layer (resolution of 80×80) and concatenating it with shallow convolutional features, which can fuse certain semantic information while preserving high spatial resolution.

4. Analysis of stage experimental results

After the Baseline method, the addition of lighting adaptive feature enhancement strategy, and the small object detection head enhancement module, the results are shown in table 2 and table 3:

Table 2. Output results of different methods

Method Name	Baseline		Add SE module		SE module+small object detection head	
Classes	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
person	18.428	26.100	18.743	30.965	19.109	28.958
rider	11.577	19.184	8.418	22.857	7.679	15.510
car	50.233	56.106	52.343	59.216	52.448	59.058
truck	23.349	31.792	22.143	38.585	22.781	33.396
bus	21.058	34.082	20.209	38.571	22.597	38.980
train	37.095	42.826	30.120	36.957	32.041	36.739
motorcycle	7.857	17.797	12.910	28.983	10.679	21.864
bicycle	8.391	16.618	6.915	14.853	7.568	11.471

Table 3. Various target AP values and mAP values for three experiments

Experiment Name	mAP	Small AP	Medium AP	Large AP
EXP1	0.2227	0.064	0.225	0.432
EXP2	0.2147	0.077	0.248	0.319
EXP3	0.2186	0.082	0.227	0.358

The above calculation results indicate that the above changes have little impact on the overall accuracy, and even have some reduction. However, looking at individual indicators, the AP value of rider gradually deteriorates with the change effect, while the AP value of car class increases

significantly by about 2%. Moreover, the mAP value of the small target has increased, by approximately 1.8% compared to the baseline results.

Based on the above results and the SE attention mechanism module added earlier, the problem is likely to lie in the SE attention mechanism module. According to Zhang Haining's research, this attention mechanism module has the drawback of only modeling channel information and ignoring the importance of spatial features in images. Therefore, in processing spatial features in images, the effect may not be as good as specialized spatial attention mechanisms. However, in this dataset, except for objects such as cars and trains that are mainly concentrated on highways, most objects are concentrated at the edges of the image, resulting in poor recognition performance. Moreover, the SE module changes the distribution of multiple channels, and in multi-objective scenarios, the channels relied upon by different target features may vary. The process of SE is globally shared, so using SE modules alone can easily lead to a situation of trade-off.

Secondly, after examining the dataset, it was found that for car like objects, ACDC dataset developers did not label some objects that were not very obvious but could be distinguished by the human eye, but these objects were recognized by the algorithm. These situations have caused a certain degree of decrease in recognition accuracy. As shown in Figure 6, this article used the Fiftyone open-source dataset annotation tool to check and found a large number of foreground objects visible to the naked eye but not labeled as cars on the front and left sides [11].



(a) Detection Results



(b) Actual annotation

Figure 6. Detection results and actual annotation (picture credit: original)

Through the intuitive output of the detection result image below, it is found that the missed detection situation is quite obvious, mainly concentrated in the edges, corners and other situations of the screen, and the algorithm is prone to ignore objects in these areas. As shown in Figure 7, people on the right side of the image with high contrast to foreground objects and more occlusion are easily recognized by the human eye, but the algorithm did not recognize them.

On the contrary, for small targets, the detection results have actually improved. As shown in Figure 8, all vehicles in the extreme distance have been successfully detected and located. From the detection box results, the effect is very good.

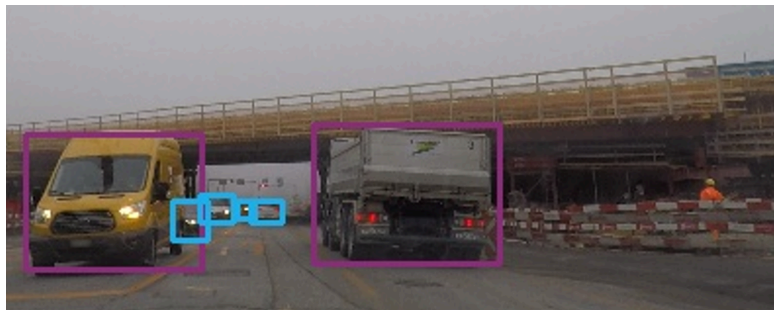


Figure 7. Schematic diagram of the missing people class on the right side of a certain image in the validation set (picture credit: original)

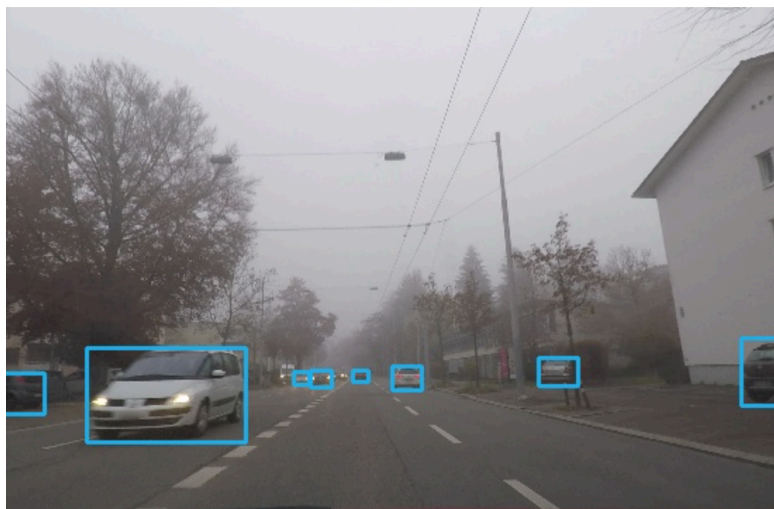


Figure 8. Verification of the detection effect of a small target car class in the validation set (picture credit: original)

5. Conclusion

From the above results, it can be seen that the above modifications did not significantly improve the performance on the ACDC dataset. The baseline method of this project only achieved an average precision of 22.27% through direct training, while the modified method showed a decrease in mAP, achieving a result of 21.86%. However, in some dimensions, the modified method also showed improvements: the accuracy of small object detection increased from 6.4% of the Baseline method to 8.2%, and the improvement in small object detection was effective. And the target for medium volume has increased from 22.5% to 24.8%. However, with the addition of attention modules, the proportion of large targets decreased from 43.2% to 31.9%.

The algorithm of Benchmarks on the official website of ACDC dataset shows that the Semi superior Object Detection for Domain Adaptation algorithm achieves 75.2% AP, while the MIC: Masked Image Consistency for Context Enhanced Domain Adaptation algorithm released in CVPR2023 achieves 46.57%, which is roughly equivalent to the AP values trained using COCO dataset mentioned by YOLOX developers in the literature.

After the above experiments, this article believes that is an improvement based on not establishing a separate image enhancement algorithm. The next improvement point is to increase the SE channel attention module, attempt to enhance the expression of image features under multi-

dimensional adverse weather conditions, and conduct experiments to observe whether the effect is improved. If the performance improvement is not significant, try using other channel attention modules for testing.

References

- [1] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. arXiv. <https://doi.org/10.48550/arXiv.2107.08430>
- [2] Megvii-BaseDetection. (2021, July 20). YOLOX: Exceeding YOLO series in 2021. *GitHub*. <https://github.com/Megvii-BaseDetection/YOLOX>
- [3] Shen, R., & Zhang, J. (2026). A review of YOLO algorithm for autonomous driving object-detection technology under adverse weather. *Journal of Computer Science and Exploration*. Advance online publication. <https://link.cnki.net/urlid/11.5602.TP.20260528.1408.006>
- [4] Wang, T., Tao, G. P., Lu, W. L., et al. (2024). Restoring vision in hazy weather with hierarchical contrastive learning. *Pattern Recognition*, 145, 109956. <https://doi.org/10.1016/j.patcog.2023.109956>
- [5] Zhang, H. (2025). Research on pedestrian-vehicle detection in blurred scenarios based on YOLO algorithm. Dalian Jiaotong University. <https://doi.org/10.26990/d.cnki.gsltc.2025.001017>
- [6] Aloufi, N., Alnori, A., & Basuhail, A. (2024). Enhancing autonomous vehicle perception in adverse weather: A multi objectives model for integrated weather classification and object detection. *Electronics*, 13(15), 3063.
- [7] Luo, K., Shen, Z., Weng, Q., et al. (2026). Robust-YOLOX-s traffic sign detection algorithm for complex weather. *Computer Simulation*, 43(4), 178-184.
- [8] Sakaridis, C., Dai, D., & Van Gool, L. (2021). ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [9] The MathWorks, Inc. (n.d.). Getting started with YOLOX for object detection. *MathWorks*. Retrieved July 24, 2026, from <https://www.mathworks.com/help/vision/ug/getting-started-with-yolox-object-detection.html>
- [10] Yang, R., Chen, L., Deng, X., et al. (2026). Research on optimization of traffic light visual detection for intelligent transportation. *Audio-Video Engineering and Application*, 50(4), 115-117. <https://doi.org/10.16311/j.audioe.2026.04.037>
- [11] Moore, B. E., & Corso, J. J. (2020). FiftyOne. *GitHub*. <https://github.com/voxel51/fiftyone>