electronics

MDPI

Article

Research on a Lightweight Object Detection Method for AGV Visual Perception Under Low-Visibility Conditions

Tao Wei [1], Huanwu Zhan [2], Shuwan Cui [3], Guiyou Zhou [3], Shibing Cai [1,*], Yilong Li [1] and Shaodong Zheng [1]

[1] Guangxi Nanguo Copper Industry Co., Ltd., Chongzuo 532103, China
[2] Nandan Nanfang Nonferrous Metals Co., Ltd., Hechi 547204, China
[3] School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China
* Correspondence: z145111705611@163.com; Tel.: +86-187-7656-5135

Abstract

With the development of intelligent manufacturing and smart logistics, automated guided vehicles (AGVs) are gradually expanding from traditional indoor environments to factory roads, logistics parks, industrial parks, and other outdoor road environments. However, low-visibility conditions, such as fog, rain, snow, and sandstorms, can degrade image quality, weaken object boundaries, and reduce the accuracy of visual perception systems. To address these challenges, an improved lightweight object detection model based on YOLOv11, named DMSM-YOLO, is proposed. To handle blurred boundaries, weak target features, and background interference in low-visibility images, the model is optimized from feature extraction, feature fusion, and prediction refinement. First, EPConv is designed to enhance directional contour interaction and improve blurred-boundary representation. Second, C3k2-MDFI is constructed to strengthen multi-scale weak-target representation. Third, MLCA is introduced to recalibrate low-contrast fused features by combining local spatial details and global contextual information. Finally, SMDetect is developed to improve localization accuracy and confidence estimation for blurred and occluded objects. Experimental results show that the proposed model achieves mAP@0.5 values of 54.71% and 71.67% on the DAWN and RTTS datasets, respectively. Compared with YOLOv11n, the mAP@0.5 is improved by 5.36 and 3.26 percentage points, respectively, while the model maintains a compact size of 2.56 M parameters and 7.1 GFLOPs. Additional experiments further verify the effectiveness of the proposed method under low-light conditions. The proposed model improves detection accuracy and robustness under low-visibility conditions while maintaining lightweight characteristics, demonstrating its potential for AGV visual perception.

Keywords: AGV; low-visibility conditions; lightweight object detection; YOLOv11

Academic Editor: John A. Kalomiros

Received: 9 August 2026
Revised: 16 September 2026
Accepted: 17 September 2026
Published: 19 September 2026

1. Introduction

Driven by intelligent manufacturing and smart logistics, AGVs are increasingly used in manufacturing plants and logistics systems [1]. Their application scenarios are gradually extending from traditional indoor environments to factory roads, logistics parks, industrial parks, and other outdoor road environments. In these outdoor scenarios, AGVs need to detect dynamic obstacles such as pedestrians, vehicles, bicycles, and trucks while dealing with perception interference caused by complex backgrounds and changing weather conditions. In particular, under low-visibility conditions such as fog, rain, snow, and sandstorms,

images captured by onboard cameras often suffer from reduced contrast, blurred boundaries, and loss of texture details, making distant, small, and occluded objects more difficult to detect accurately. These problems increase missed detections and false alarms, thereby affecting obstacle-avoidance decisions and operational safety [2,3]. Therefore, object detection models for outdoor AGVs should not only adapt to low-visibility environments but also meet the lightweight and low-computation requirements of onboard edge devices.

With the development of intelligent perception technology, object detection has been widely applied in computer vision and intelligent transportation. According to sensing modalities, existing perception methods can be roughly divided into LiDAR-based and vision-based approaches. Although LiDAR-based methods can provide accurate spatial information, their high cost and deployment complexity limit their large-scale practical application. In contrast, camera-based visual detection methods are less costly and more flexible to deploy, making them an important technical route for visual perception. Representative visual detection frameworks include one-stage detectors such as YOLO [4] and two-stage detectors such as R-CNN [5]. Two-stage detectors generally achieve high detection accuracy but suffer from relatively slow inference speed, whereas YOLO-based detectors have attracted extensive attention because of their real-time performance and deployment efficiency. To meet the requirements of real-time road perception, researchers have proposed various improved YOLO methods for obstacle detection and traffic-object recognition tasks. Wang et al. [6] proposed YOLO-OD for vision-assisted obstacle detection, improving the practicality of obstacle perception. Cui et al. [7] proposed DAN-YOLO for adverse-weather autonomous driving scenarios, achieving a good balance between detection accuracy and inference speed. Cai et al. [8] proposed MFMAM-YOLO for pole-like obstacle detection in complex environments, improving the detection capability for slender road objects. Sun et al. [9] proposed an improved YOLOv5 model for pedestrian and vehicle detection, enhancing detection accuracy in road scenes. These studies show that YOLO-based detectors have good application potential in practical perception tasks. However, most methods are mainly designed for general traffic scenarios or specific obstacle categories, and their adaptability to outdoor AGV visual perception under low-visibility conditions remains insufficiently investigated.

In recent years, AGV-oriented perception and obstacle-avoidance methods have also attracted increasing attention. Yang et al. [10] proposed a global-vision-based multi-AGV tracking system that combines April Tag and an Extended Kalman Filter to improve localization and scheduling efficiency in intelligent warehouses. Yang et al. [11] designed an improved YOLOv5n6 model for shelter identification in shelter-transporting AGVs, improving detection accuracy with only a slight increase in detection time. Zhang et al. [12] proposed an improved YOLOv5s detector for complex traffic environments, achieving real-time AGV-related target detection. Tan et al. [13] proposed a multi-sensor fusion obstacle-avoidance method based on improved YOLOv8 to enhance autonomous obstacle avoidance for AGVs. Cai et al. [14] proposed LSOD-YOLO for AGV perception systems, reducing model parameters while improving detection accuracy. In addition, Tai et al. [15] proposed an optimized YOLOv3-based dynamic recognition algorithm to reduce model complexity and computational cost, while Wu et al. [16] explored multi-sensor perception optimization for AGV navigation. Although these methods improve AGV perception performance in terms of localization, detection accuracy, inference efficiency, or model complexity, most of them are still evaluated under normal imaging conditions or relatively structured environments. For outdoor AGVs operating in low-visibility road scenes, degraded images introduce blurred contours, weak textures, low contrast, and weather-induced noise, making it difficult for lightweight detectors to extract discriminative features from distant, blurred, or partially occluded objects. Therefore, it is still necessary to develop

a lightweight object detection model that can enhance degraded visual features while maintaining low computational cost for edge deployment.

To address the above challenges, this study proposes DMSM-YOLO, a lightweight object detection model based on YOLOv11n for outdoor AGV visual perception under low-visibility conditions. To address feature degradation caused by low-visibility imaging, targeted improvements are introduced at the feature extraction, feature fusion, and prediction stages. Specifically, an Enhanced Pinwheel Convolution (EPConv) is proposed to improve blurred-boundary representation. Different from the independent directional branches in PConv, EPConv introduces diagonal residual interaction before feature fusion, enabling complementary directional information to compensate for weakened contour responses. A Multi-scale Degradation-aware Feature Interaction (MDFI) block is designed and embedded into C3k2 to construct C3k2-MDFI. By combining multi-scale dilated spatial modeling with multiplicative feature interaction, it strengthens the representation of weak, small, and scale-varying targets under low-visibility conditions. A stage-specific feature refinement scheme is constructed in the neck and detection head. MLCA is selectively introduced after multi-scale feature fusion, while SimAM is embedded into the regression and classification branches to construct SMDetect, extending degraded-feature enhancement from feature extraction to feature fusion and final prediction. Experiments on the DAWN and RTTS datasets show that DMSM-YOLO improves mAP@0.5 by 5.36 and 3.26 percentage points over YOLOv11n, respectively. Additional experiments on the BDD100K low-light subset further achieve a 3.12 percentage-point improvement in mAP@0.5. These results indicate that DMSM-YOLO can effectively improve object detection performance under different low-visibility conditions while maintaining its lightweight characteristics, making it more suitable for outdoor AGV object detection in degraded visual environments.

2. Method

2.1. Baseline Model and Improvements

YOLOv11 [17], developed on the YOLOv8 framework [18], achieves a favorable balance between parameter size and computational efficiency compared to YOLOv9 [19] and YOLOv10 [20]. YOLOv11n balances performance, cost, and speed, suiting resource-constrained edge devices, thus serving as the baseline, with its overarching architecture presented in Figure 1.

The architecture comprises a backbone, neck, and detection head. The backbone integrates convolutions, C3k2 blocks, SPPF, and C2PSA. The C3k2 module succeeds YOLOv8's C2f to improve feature extraction, while the post-SPPF C2PSA enhances representation via position-sensitive attention. The neck fuses P3/8, P4/16, and P5/32 scales via top-down FPN and bottom-up PAN using C3k2 blocks. The detection head employs an anchor-free decoupled layout for multi-scale end-to-end predictions.

Based on YOLOv11n, four improvements are introduced into the backbone, neck, and detection head to enhance object detection under low-visibility conditions. First, EPConv is used to replace part of the standard convolutional operations in the backbone, thereby strengthening multi-directional feature extraction and improving the representation of blurred object boundaries. Second, the C3k2-MDFI module is constructed by embedding the proposed MDFI Blocks into the C3k2 structure, which improves the extraction of weak and small targets. Third, MLCA is embedded after neck feature fusion to enhance the interaction between local spatial features and global channel information. Finally, SMDetect is designed by introducing SimAM into the classification and regression branches of the detection head, improving the localization and confidence prediction of degraded targets. The overall architecture of DMSM-YOLO is shown in Figure 2.

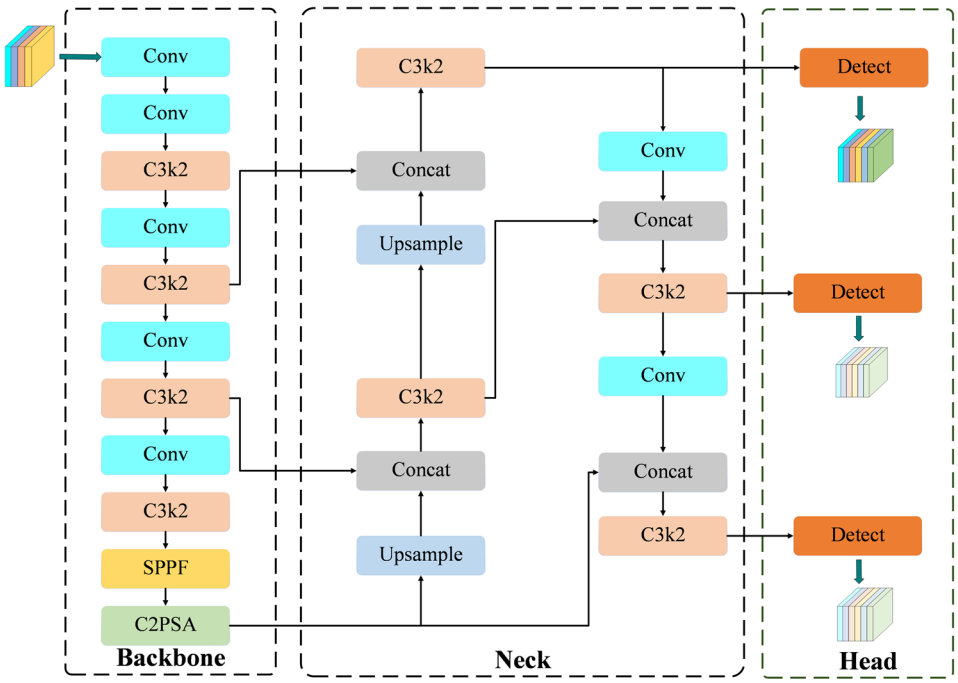


Figure 1. Overall structure of YOLOv11.

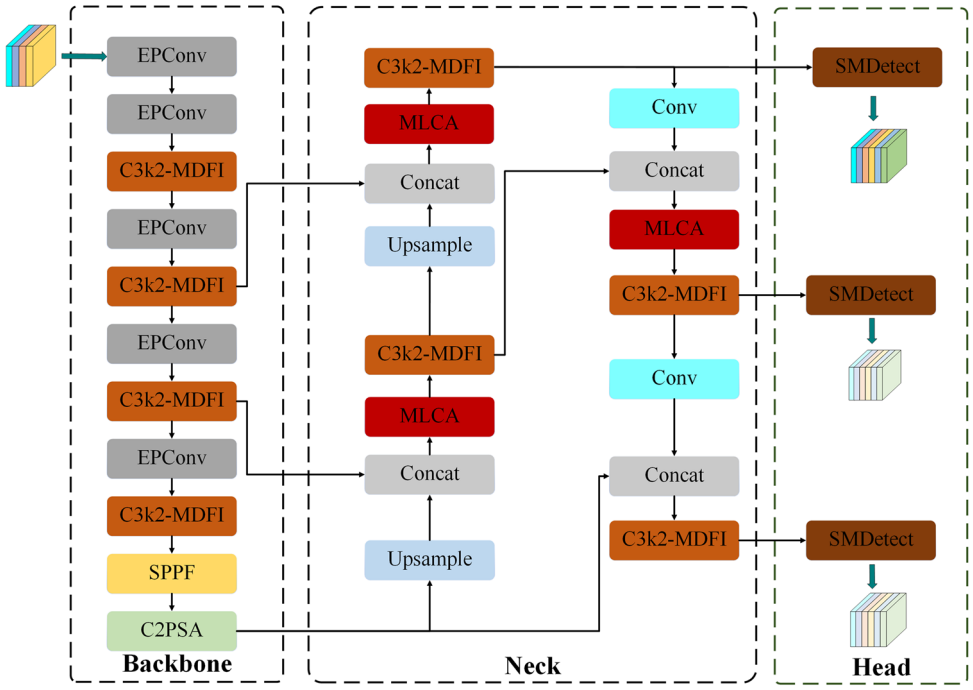
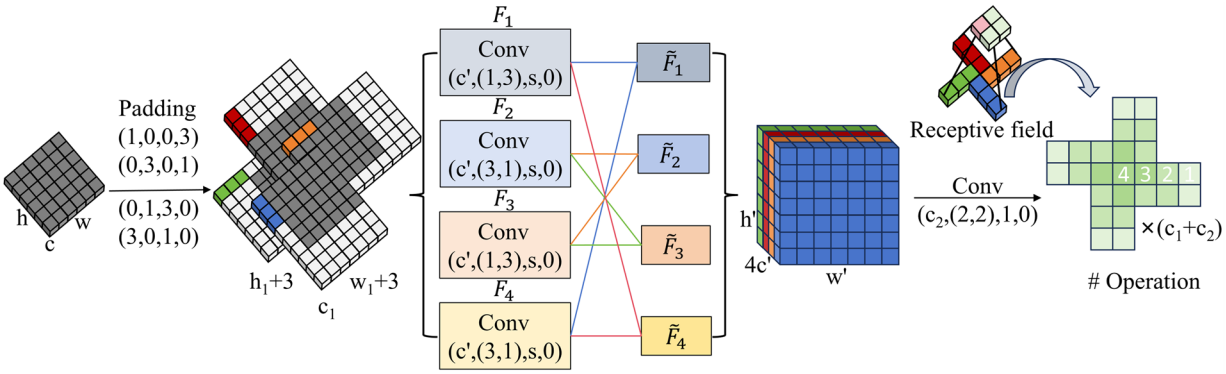


Figure 2. Overall architecture of DMSM-YOLO.

2.2. EPConv Design

Under low-visibility conditions, weather degradation such as fog, rain, and snow weakens object contours, reduces boundary contrast, and introduces direction-dependent interference. Conventional convolution samples features in a symmetric local window, which is effective for general texture extraction but is insufficient for modeling directional contour cues in degraded images. PConv [21] improves directional perception by using four asymmetric padding strategies to extract features from different spatial directions. However, the directional branches in the original PConv are processed independently and

are fused only by channel concatenation. This independent-then-concatenate strategy limits information exchange among directional responses, especially when object boundaries are blurred or partially missing. To address this limitation, this study proposes Enhanced Pinwheel Convolution (EPConv), as shown in Figure 3. Different from the original PConv, EPConv introduces a diagonal residual interaction strategy before channel concatenation. Specifically, the four directional branches are divided into two diagonal complementary pairs, namely branches (1, 4) and branches (2, 3). For each pair, the response of one branch is enhanced by a weighted residual response from its diagonal counterpart. In this way, weak directional features can be compensated by complementary diagonal information before final feature fusion.



Padding(left, right, top, bottom); Conv= SiLU(BN(Conv2d(c_{out}, k, s, p)))

Figure 3. Principle of the EPConv architecture.

The forward propagation of EPConv is formulated as follows:

$$F_i = P_i(X),\ i = 1,\ 2,\ 3,\ 4 \tag{1}$$

$$\widetilde{F}_1 = F_1 + \beta F_4,\ \widetilde{F}_4 = F_4 + \beta F_1 \tag{2}$$

$$\widetilde{F}_2 = F_2 + \beta F_3,\ \widetilde{F}_3 = F_3 + \beta F_2 \tag{3}$$

$$Y = C\left(Concat\left(\widetilde{F}_1,\ \widetilde{F}_2,\ \widetilde{F}_3,\ \widetilde{F}_4\right)\right) \tag{4}$$

where X denotes the input feature map, P_i represents the i-th asymmetric directional branch, and F_i is the corresponding directional feature. where β is the diagonal fusion weight, and C denotes the channel fusion operation.

An appropriate value of β is expected to balance the preservation of directional feature specificity and the introduction of complementary diagonal information. When β = 0, no cross-branch compensation is performed, and the structure is equivalent to the original PConv. In contrast, an excessively large β may introduce excessive cross-branch information and weaken the original directional asymmetry. By enabling diagonal information compensation before feature fusion, EPConv reduces the loss of weak contour cues caused by independent branch extraction. This helps the backbone obtain more stable boundary-aware features for blurred, low-contrast, and partially occluded objects, while the fixed fusion weight keeps the module lightweight.

2.3. C3k2-MDFI Design

Under low-visibility conditions, object features are easily degraded by blurred contours, weak texture responses, reduced contrast, and scale variation. These degradation factors are more obvious for small, distant, and partially occluded objects, making it dif-

ficult for lightweight detectors to extract discriminative features. In YOLOv11n, C3k2 is used as an efficient feature extraction module to improve gradient flow and feature reuse. Its internal Bottleneck units can effectively extract local features under normal imaging conditions. However, in low-visibility scenes, object contours are often blurred, texture details are weakened, and targets show obvious scale variation.

To overcome these limitations, this study proposes a Multi-scale Degradation-aware Feature Interaction (MDFI) Block. The internal structure of MDFI is shown in Figure 4. MDFI first constructs a multi-scale degradation-aware spatial mapping to enhance its adaptability to degraded images. Specifically, three lightweight spatial branches with different dilation rates are used to capture contextual information from different receptive fields. The aggregated multi-scale features are then transformed into two independent feature mappings through lightweight channel mappings. Based on the multiplicative feature-interaction strategy [22], the two feature mappings are further interacted in an element-wise manner to enhance nonlinear feature representation. After spatial refinement, a residual connection is introduced to preserve original feature information and stabilize feature propagation. The multi-scale mapping is defined as

$$M(X) = \frac{1}{S}\sum_{s=1}^{S} D_s(X),\ S = 3 \tag{5}$$

where X is the input feature map, $M(X)$ is the multi-scale aggregated feature, and $D_s(X)$ denotes the output of the s-th spatial branch. The three branches adopt dilation rates of 1, 2, and 3, corresponding to effective receptive fields of 3×3, 5×5, and 7×7, respectively.

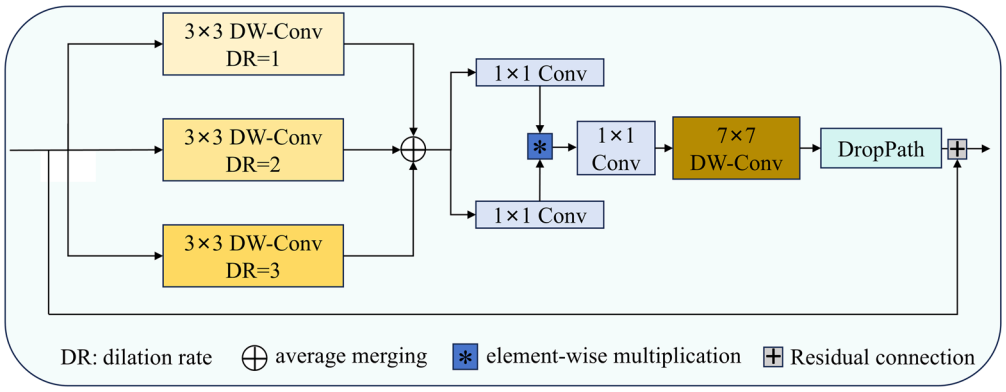


Figure 4. Principle of the MDFI architecture.

After multi-scale contextual aggregation, MDFI performs multiplicative nonlinear feature interaction:

$$U = P_1(M(X)),\ V = P_2(M(X)) \tag{6}$$

$$F = \delta(U) \otimes V \tag{7}$$

$$Y = X + T(P_o(F)) \tag{8}$$

where U and V are two independently mapped features, F is the interacted feature, and Y is the output of the MDFI Block. P_1, P_2, and P_o denote 1×1 convolutional layers, δ is the ReLU activation function, $\otimes$ denotes element-wise multiplication, and T represents spatial contextual refinement comprising a 7×7 depthwise convolution and DropPath regularization.

As shown in Figure 5, the proposed C3k2-MDFI module is obtained by embedding MDFI Blocks into the original C3k2 structure. This design retains the efficient feature reuse framework of C3k2 while introducing multi-scale degradation-aware spatial modeling and

nonlinear feature interaction. Compared with the original C3k2 module, C3k2-MDFI can better enhance weak-target features, low-contrast regions, and blurred object boundaries under low-visibility conditions. Meanwhile, the lightweight design of MDFI enables feature representation to be improved without significantly increasing model complexity.

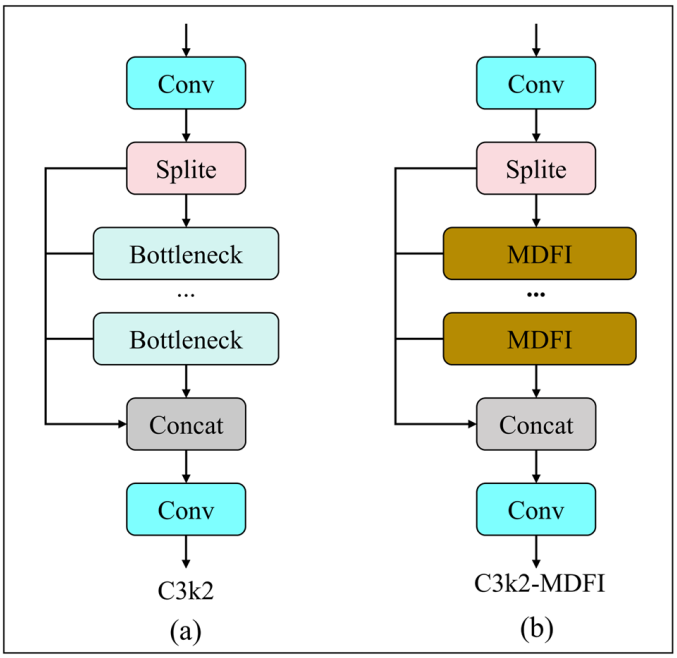


Figure 5. Structure of the C3k2-MDFI module. (**a**) Original module; (**b**) Improved module.

2.4. MLCA Mechanism

Under low-visibility conditions, reduced contrast and weather-induced degradation weaken target responses and increase background interference. Although the neck network fuses multi-scale features, the fused features may still contain redundant background information in low-contrast regions. To improve feature discrimination after feature fusion, the Mixed Local Channel Attention (MLCA) module proposed by Wan et al. [23] is introduced into the neck, as shown in Figure 6. MLCA jointly models local spatial information and global channel context. It first obtains local descriptors through local average pooling and then extracts global descriptors from the local representation. One-dimensional convolution is used to model cross-channel interactions without dimensionality reduction.

In the proposed network architecture, the insertion locations of MLCA are strictly governed by the trade-off between spatial resolution and computational efficiency. Specifically, MLCA is explicitly positioned after the first three multi-scale feature-fusion operations in the neck. This strategic placement is driven by the fact that shallow and mid-level fusion maps retain rich, high-resolution spatial details, making the local-global attention calibration highly effective in recovering weak object responses against weather interference. Conversely, forcing MLCA into deeper fusion layers—where spatial resolution is highly compressed—yields marginal calibration benefits while introducing redundant computational overhead. This selective insertion scheme balances feature enhancement and computational overhead while preserving the lightweight characteristics of the model.

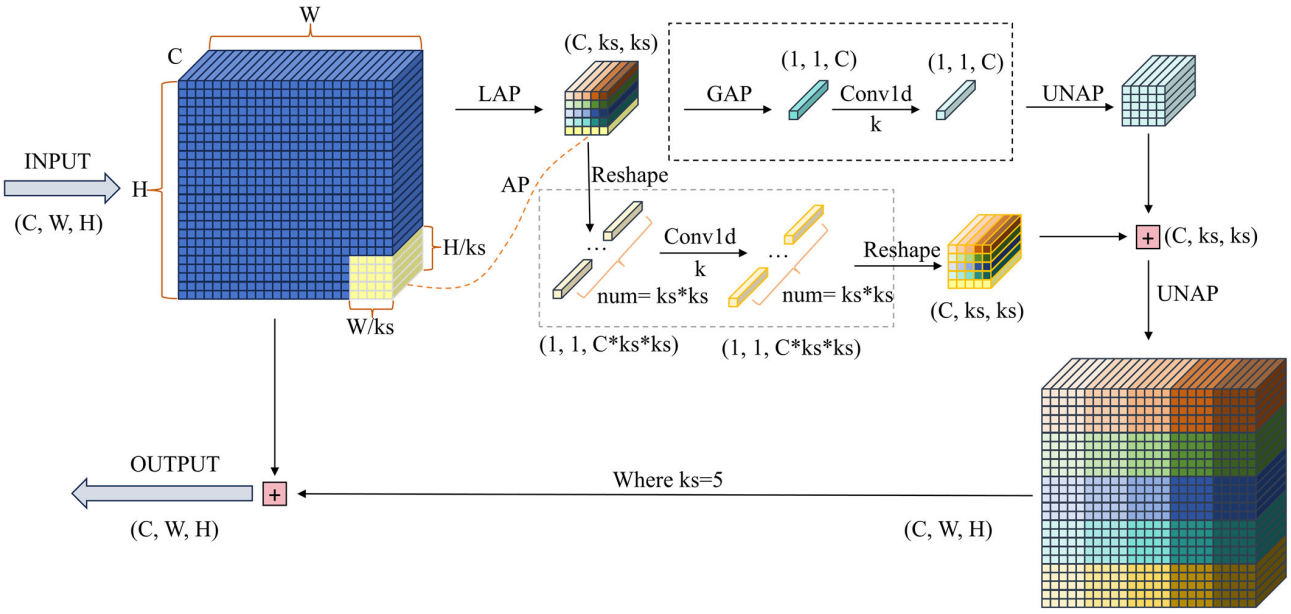


Figure 6. Principle of the MLCA architecture.

2.5. SMDetect

In low-visibility scenes, target textures are often weakened, object boundaries become blurred, and background noise may interfere with the final prediction. Although the original YOLOv11 detection head adopts a decoupled structure to separately perform classification and bounding box regression, the feature responses before prediction may still be insufficient for blurred, occluded, and low-contrast objects. This can reduce classification confidence and weaken localization accuracy.

To enhance the prediction ability of the detection head under degraded imaging conditions, this study introduces the SimAM attention mechanism [24], whose structure is shown in Figure 7. SimAM is a parameter-free attention mechanism that estimates neuron importance through an energy function and generates three-dimensional attention weights without introducing additional learnable parameters. Its attention weight can be calculated as

$$A = \sigma\left(\frac{(X-\mu)^2}{4(\sigma^2+\lambda)} + 0.5\right) \tag{9}$$

where X denotes the input feature map, μ and σ^2 represent the mean and variance of the feature map, respectively, λ is a smoothing coefficient, and σ denotes the sigmoid activation function. The recalibrated feature is then obtained by

$$Y = X \otimes A \tag{10}$$

where $\otimes$ denotes element-wise multiplication.

Based on SimAM, an improved detection head named SMDetect is designed by embedding SimAM into both the classification and regression branches, as shown in Figure 8. In SMDetect, SimAM is placed after the feature convolution layers and before the final prediction convolution in both branches. In the regression branch, SimAM enhances boundary-related and localization-sensitive responses before bounding box prediction, which is beneficial for blurred and partially occluded objects. In the classification branch, SimAM strengthens discriminative target responses and suppresses background interference before category prediction, thereby improving confidence estimation for low-contrast objects. Since SimAM does not introduce additional learnable parameters, SMDetect im-

proves the detection head's feature recalibration ability while preserving the lightweight property of the model.

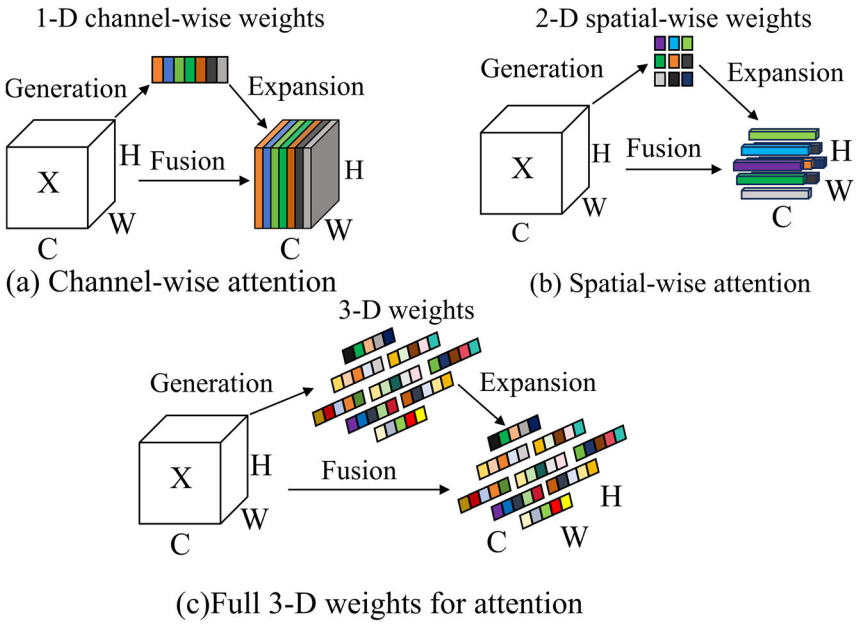


Figure 7. The SimAM architecture.

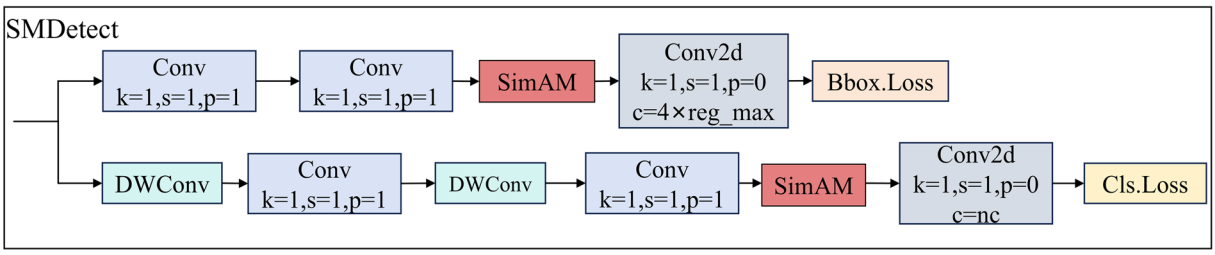


Figure 8. Architecture and working principle of SMDetect.

3. Experiments and Results Analysis

3.1. Dataset

In this study, two publicly available datasets, DAWN [25] and RTTS [26], were adopted for experimental evaluation. These datasets contain representative low-visibility traffic scenes and are used to evaluate the proposed model under mixed adverse-weather and fog-dominated conditions. The DAWN dataset consists of 1, 027 real-world images captured under various adverse weather conditions, including fog, rain, snow, and dust, and contains object categories related to road perception, such as cars, persons, buses, trucks, motorcycles, and bicycles. The RTTS dataset contains 4, 322 real-world foggy images collected from diverse scenes, covering different viewing angles and reduced-visibility environments with significant fog interference. For both datasets, the images were randomly divided into training, validation, and testing sets at a ratio of 7:2:1. Although these public datasets differ from actual outdoor AGV environments in terms of scene structure and object distributions, they cover typical low-visibility conditions that are also relevant to outdoor AGV visual perception. In particular, weather-induced degradation such as reduced contrast, blurred object boundaries, and weakened target features is consistent with the visual challenges considered in this study. Therefore, DAWN and RTTS are used to evaluate the detection robustness of the proposed model under representative low-visibility conditions. Detailed dataset configurations are presented in Table 1.

Table 1. Statistics of the DAWN and RTTS datasets.

Dataset	Number of Images	Classes
DAWN	train:718, val:206, test: 103	car (82.27%), person (6.07%), bus (2.05%), truck (8.23%), motorcycle + bicycle (1.38%)
RTTS	train:3026, val:864, test: 432	person (27.94%), car (61.26%), bus (6.07%), motorbike (3.01%), bicycle (1.72%)

3.2. Experimental Parameters

To evaluate both detection accuracy and lightweight performance, this study adopts Precision, Recall, mAP@0.5, mAP@0.5:0.95, Parameters, GFLOPs, and FPS as evaluation metrics. Precision and Recall are used to measure false detections and missed detections, respectively. Their calculation formulas are given in Equations (11) and (12):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The mean average precision (mAP) is used to evaluate the overall detection performance across all categories and is calculated according to Equation (13):

$$\text{mAP} = \frac{1}{\text{n}} \sum_{\text{k}=1}^{\text{n}} \text{AP}_{\text{k}} \tag{13}$$

All experiments were performed on a Windows 11 workstation with an Intel Core i9-14900K CPU and an NVIDIA GeForce RTX 4090D GPU with 24 GB of VRAM. The implementation was based on Python 3.10, PyTorch 2.3.1, and CUDA 12.1. For fair comparison, all models were trained from epoch 0 under the same training protocol. The detailed training settings are listed in Table 2.

Table 2. Training parameter settings.

Parameters	Input Size	Epochs	Batch Size	Optimizer	Learning Rate	Momentum	Weight Decay
Value	640 × 640	200	32	SGD	0.01	0.937	0.0005

3.3. EPConv Parameter Analysis

To determine the optimal cross-branch fusion weight β in EPConv, experiments were conducted with different β values. β controls the contribution of diagonal branch information. When $\beta = 0$, the diagonal cross-branch interaction is disabled, equivalent to the original PConv. As shown in Figure 9a, the model achieves the highest detection accuracy at $\beta = 0.25$. Appropriate diagonal interaction improves detection performance, while excessive β weakens the asymmetric directional sensitivity and causes feature interference. To further visualize the effect of β, Figure 9b presents Grad-CAM heatmaps: at $\beta = 0$, responses are scattered and boundaries are blurred; at $\beta = 0.25$, responses concentrate on object contours; at $\beta \geq 0.5$, background activation increases significantly. Based on both quantitative and visual analysis, $\beta = 0.25$ is selected as the optimal fusion weight for EPConv.

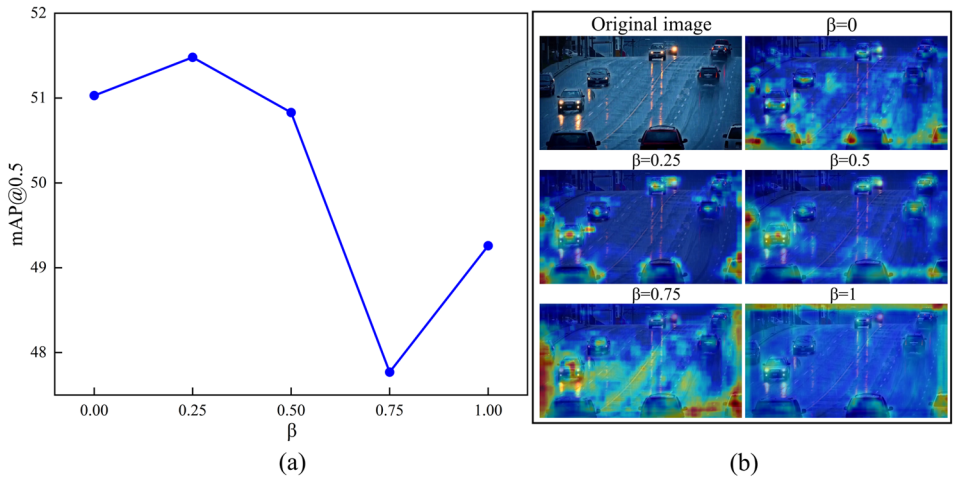


Figure 9. Performance comparison under different fusion weights β. (**a**) mAP@0.5 of different β values. (**b**) Feature heatmap visualization under different β values.

3.4. Comparative Experiments of Different Attention Mechanisms

To validate the effectiveness of MLCA, comparative experiments on different attention mechanisms (ECA [27], SE [28], CA [29], CBAM [30]) were conducted on the DAWN dataset. As shown in Table 3, MLCA achieves the highest mAP@0.5 of 51.02%, which is 0.22% higher than ECA, the second best at 50.80%. In terms of precision, MLCA reaches 68.70%, also surpassing other methods. For recall, MLCA achieves 46.84%, while CA yields the highest value of 48.26%. Regarding model complexity, all methods have approximately 2.58 M parameters, and MLCA's GFLOPs is 6.4 G, only 0.1 G higher than others. Overall, MLCA achieves the highest mAP@0.5 and Precision while maintaining comparable model complexity.

Table 3. Comparative experiments of different attention mechanisms.

Mechanisms	Precision/%	Recall/%	mAP@0.5/%	Parameters/M	GFLOPs/G
ECA	67.26	46.12	50.80	2.58	6.3
SE	63.34	47.62	50.11	2.59	6.3
CA	58.76	48.26	50.51	2.59	6.3
CBAM	67.14	39.81	49.87	2.58	6.3
MLCA	68.70	46.84	51.02	2.58	6.4

3.5. Ablation Experiments

Ablation experiments were conducted on the DAWN dataset to evaluate the contribution of each component in DMSM-YOLO. Table 4 reports the ablation results of EPConv, C3k2-MDFI, MLCA, and SMDetect under the same training settings. Compared with the baseline YOLOv11n, introducing EPConv alone improves mAP@0.5 by 2.13%, indicating that diagonal cross-branch interaction enhances directional feature representation under low-visibility conditions. When C3k2-MDFI is used alone, mAP@0.5 increases by 3.29 percentage points, indicating the effectiveness of multi-scale spatial modeling and multiplicative feature interaction for weak-target representation. In addition, MLCA improves mAP@0.5 by 1.67%, while SMDetect improves mAP@0.5 by 1.96%, demonstrating that local-global contextual fusion and detection-head saliency recalibration both contribute to low-visibility object detection. When all four components are integrated, DMSM-YOLO achieves 54.71% mAP@0.5, representing a 5.36% improvement over YOLOv11n. Meanwhile, Precision and Recall reach 76.88% and 46.85%, respectively. In terms of per-class performance, the most significant improvements are observed on bus and motorcycle +

bicycle, with improvements of 10.81% and 9.17% respectively, indicating that the model is particularly effective for small and texture-weak targets under low-visibility conditions. In terms of model complexity, the number of parameters slightly decreases from 2.58 M to 2.56 M, while FLOPs increase from 6.3 G to 7.1 G. These results indicate that the proposed method improves detection accuracy while maintaining a lightweight model structure. Although the sum of the individual improvements is higher than the improvement of the integrated model, the final model achieves the best overall performance, suggesting that the proposed components work collaboratively to enhance detection robustness under low-visibility conditions.

Table 4. Ablation results on the DAWN dataset.

EPConv	C3k2-MDFI	MLCA	SMDetect	Precision/%	Recall/%	mAP@0.5/%	Parameters/M	GFLOPs/G	A/%	B/%	C/%	D/%	E/%
				63.96	44.35	49.35	2.58	6.3	52.94	39.53	86.11	33.03	35.14
√				61.40	47.06	51.48	2.52	6.9	56.11	39.81	86.63	32.45	42.39
	√			76.26	45.22	52.64	2.62	6.4	54.52	38.81	86.38	40.88	42.59
		√		68.70	46.84	51.02	2.58	6.4	51.14	44.47	86.97	35.29	37.25
			√	71.37	45.60	51.31	2.58	6.3	54.05	42.51	86.85	33.94	39.20
√	√			72.89	46.60	53.45	2.56	7.0	53.11	41.33	86.64	37.10	49.06
√	√	√		68.85	49.79	54.27	2.56	7.1	56.29	41.99	87.05	41.06	44.98
√	√	√	√	76.88	46.85	54.71	2.56	7.1	55.36	43.23	86.79	43.84	44.31

Note: A: truck; B: person; C: car; D: bus; E: motorcycle + bicycle. √ indicates that the corresponding module was used.

Ablation experiments were further conducted on the RTTS dataset to verify the effectiveness of each component in foggy scenarios. Consistent with the trends on the DAWN dataset, each individual module contributes positively, as shown in Table 5, with C3k2-MDFI achieving the largest gain of 2.20%, followed by MLCA with 1.34%, EPConv with 1.19%, and SMDetect with 0.99%. When all four modules are integrated, mAP@0.5 reaches 71.67%, representing a 3.26% improvement over the baseline. In terms of per-class performance, the most significant improvements are observed on bicycle and motorbike, with improvements of 5.58% and 4.89% respectively, while the car class exhibits limited improvement of 1.37% due to its already high baseline AP of 82.89%. These results further validate the effectiveness and robustness of DMSM-YOLO under fog-dominated low-visibility conditions.

Table 5. Ablation results on the RTTS dataset.

EPConv	C3k2-MDFI	MLCA	SMDetect	Precision/%	Recall/%	mAP@0.5/%	Parameters/M	GFLOPs/G	A/%	B/%	C/%	D/%	E/%
				75.68	60.60	68.41	2.58	6.3	61.28	60.83	82.89	61.63	75.40
√				77.98	59.99	69.60	2.52	6.9	64.15	63.13	82.46	63.64	74.64
	√			76.71	62.56	70.61	2.62	6.4	64.22	61.98	83.93	65.34	77.57
		√		76.48	60.35	69.75	2.58	6.4	65.04	62.62	83.07	62.54	75.45
			√	71.84	62.66	69.40	2.58	6.3	64.18	61.35	83.36	62.26	75.86
√	√			75.98	64.21	71.24	2.56	7.0	64.26	63.17	84.62	66.27	77.89
√	√	√		76.52	64.36	71.41	2.56	7.1	64.52	64.21	84.99	65.39	77.95
√	√	√	√	76.59	63.62	71.67	2.56	7.1	66.86	63.29	84.26	66.52	77.41

Note: A: bicycle; B: bus; C: car; D: motorbike; E: person. √ indicates that the corresponding module was used.

3.6. Comparative Experiments

Table 6 presents the comparison results between the proposed DMSM-YOLO and mainstream YOLO models on the DAWN dataset, including YOLO26 [31], YOLOv13 [32], YOLOv12 [33], etc. As shown in Table 6, DMSM-YOLO achieves 54.71% mAP@0.5 and 33.38% mAP@0.5:0.95. Compared with the baseline YOLOv11n, mAP@0.5 and mAP@0.5:0.95 are improved by 5.36 and 3.90 percentage points, respectively, while Precision increases from 63.96% to 76.88%. These results indicate that the proposed method improves detection accuracy and reduces false detections in mixed low-visibility scenes. Compared

with the n-series models, DMSM-YOLO obtains the best mAP@0.5 performance while maintaining a compact model size. Compared with the s-series models, DMSM-YOLO surpasses most s-series models in mAP@0.5. Although its mAP@0.5 is slightly lower than YOLOv11s and YOLOv8s, DMSM-YOLO requires only 2.56 M parameters and 7.1 GFLOPs, which are much lower than those of YOLOv11s and YOLOv8s. Meanwhile, DMSM-YOLO achieves an inference speed of 214 FPS, indicating that the proposed improvements maintain high inference efficiency under the same experimental setting. Compared with the three SOTA methods (YOLO-FOD [34], YOLOv8-STE [35], Fog-YOLO11n [36]), DMSM-YOLO achieves a higher mAP@0.5 than YOLO-FOD while requiring fewer parameters and GFLOPs. However, YOLOv8-STE and Fog-YOLO11n achieve higher mAP@0.5 than DMSM-YOLO. Nevertheless, DMSM-YOLO maintains a compact parameter scale, achieves higher Precision than Fog-YOLO11n, and requires substantially fewer parameters and GFLOPs than YOLOv8-STE.

Table 6. Comparative results on the DAWN dataset.

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs/G	FPS
YOLO26n	66.85	45.42	48.99	29.68	2.38	5.2	274
YOLO26s	67.13	42.55	50.34	30.48	9.47	20.5	157
YOLOv13n	56.22	44.76	46.82	28.03	2.45	6.2	253
YOLOv13s	75.54	47.28	53.69	31.62	9.00	20.7	110
YOLOv12n	66.41	45.55	49.81	29.34	2.55	6.3	245
YOLOv12s	58.98	50.98	53.29	32.02	9.23	21.2	128
YOLOv11n	63.96	44.35	49.35	29.48	2.58	6.3	225
YOLOv11s	70.25	51.17	55.02	33.98	9.41	21.3	122
YOLOv10n	59.22	42.88	46.12	27.47	2.70	8.2	293
YOLOv10s	60.09	50.15	52.07	32.56	8.04	24.5	253
YOLOv9t	63.03	45.91	48.42	30.08	1.73	6.4	212
YOLOv9s	65.29	47.57	53.42	33.52	6.20	22.1	103
YOLOv8n	66.74	44.48	49.44	29.70	2.69	6.8	254
YOLOv8s	73.15	50.48	56.58	35.26	9.83	23.4	150
YOLO-FOD	/	/	50.0	/	6.7	16.0	124
YOLOv8-STE	/	/	68.26	/	20.3	48.3	/
Fog-YOLO11n	72.5	62.8	63.60	34.80	2.14	5.66	89
DMSM-YOLO	76.88	46.85	54.71	33.38	2.56	7.1	214

Note: Results of the SOTA methods are taken from the original publications. Differences exist among the SOTA methods in dataset settings, training data, input resolution, evaluation metrics, and speed measurement protocols. "/" denotes information or metrics not reported in the original publications.

Figure 10 compares the detection results of YOLOv11n and DMSM-YOLO on the DAWN test dataset. Four representative low-visibility scenarios are selected, including fog, snow, sandstorm, and rain. Figure 10a shows the original test images, and Figure 10b presents the corresponding ground-truth labels. As shown in Figure 10c, YOLOv11n tends to miss distant or blurred objects under severe weather degradation, especially small targets and partially occluded objects. Some predicted bounding boxes also show relatively low confidence or incomplete localization. In contrast, DMSM-YOLO detects more degraded targets in these challenging scenes, as shown in Figure 10d. The proposed method produces more stable localization results and higher confidence scores for several blurred and occluded objects. Figure 10e presents Grad-CAM visualizations, showing that DMSM-YOLO produces relatively more concentrated responses around object regions, whereas Figure 10f indicates that YOLOv11n exhibits broader activation extending into some surrounding background regions. These visual comparisons further indicate that DMSM-YOLO improves object detection robustness under mixed low-visibility conditions.

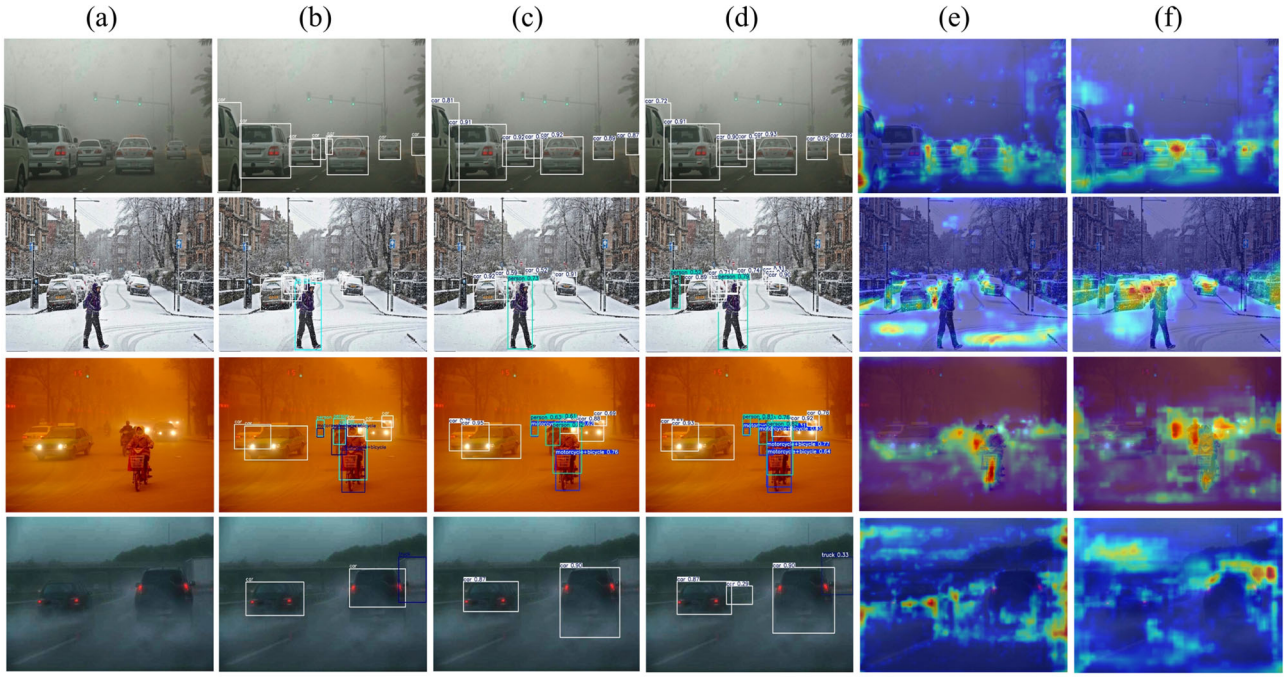


Figure 10. Detection results on the DAWN test dataset. (**a**) Original image; (**b**) ground truth; (**c**) YOLOv11n; (**d**) DMSM-YOLO; (**e**) DMSM-YOLO Grad-CAM; (**f**) YOLOv11n Grad-CAM.

Table 7 reports the comparative results on the RTTS dataset, which mainly contains real-world foggy scenes. Compared with DAWN, RTTS is used to further evaluate the robustness of the proposed method under fog-dominated low-visibility degradation. As shown in Table 7, DMSM-YOLO achieves 71.67% mAP@0.5 and 46.01% mAP@0.5:0.95. Compared with YOLOv11n, mAP@0.5 and mAP@0.5:0.95 are improved by 3.26% and 1.45%, respectively. In particular, Recall increases from 60.60% to 63.62%, indicating that the proposed method can detect more targets in foggy scenes and reduce missed detections. Compared with the n-series models, DMSM-YOLO achieves the best mAP@0.5 performance on RTTS. When compared with the s-series models, several larger models obtain slightly higher mAP@0.5, but DMSM-YOLO still maintains effective detection performance. Similarly, in comparison with recent SOTA methods, DMSM-YOLO outperforms YOLO-FOD with fewer parameters, while Fog-YOLO11n achieves higher accuracy with a lighter model, and YOLOv8-STE reaches the highest accuracy with substantially larger model size. Notably, DMSM-YOLO exhibits advantages in recall and mAP@0.5:0.95 over Fog-YOLO11n. Overall, the RTTS results further demonstrate the effectiveness of DMSM-YOLO for blurred and low-contrast object detection under foggy low-visibility conditions.

Figure 11 compares the detection results of YOLOv11n and DMSM-YOLO on the RTTS test dataset. The selected images contain typical foggy low-visibility scenes, where object contours are blurred and background contrast is significantly reduced. Figure 11a displays the original foggy test images, and Figure 11b provides the corresponding ground-truth annotations. As shown in Figure 11c, YOLOv11n misses some distant vehicles, pedestrians, and crowded targets under heavy fog interference. In several cases, the predicted boxes also show relatively low confidence because the target features are weakened by haze. In contrast, DMSM-YOLO produces more complete detection results, as shown in Figure 11d. The proposed method detects more blurred and low-contrast objects in foggy scenes and improves the confidence of several detected targets. Figure 11e presents Grad-CAM visualizations, showing that DMSM-YOLO produces more focused responses on the target regions in foggy scenes, whereas Figure 11f shows that YOLOv11n yields relatively broader activation, with some responses extending into the surrounding back-

ground. These visual results further support the effectiveness of DMSM-YOLO under fog-dominated low-visibility conditions.

Table 7. Comparative results on the RTTS dataset.

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/M	GFLOPs/G
YOLO26n	72.41	59.52	67.88	44.00	2.38	5.2
YOLO26s	79.14	61.94	70.54	46.41	9.47	20.5
YOLOv13n	75.70	60.39	68.48	43.92	2.45	6.2
YOLOv13s	76.91	63.26	70.99	46.57	9.00	20.7
YOLOv12n	73.39	61.83	68.98	44.74	2.55	6.3
YOLOv12s	79.78	63.67	72.34	47.99	9.23	21.2
YOLOv11n	75.68	60.60	68.41	44.56	2.58	6.3
YOLOv11s	80.31	62.53	71.93	47.65	9.41	21.3
YOLOv10n	75.31	60.79	68.25	44.10	2.70	8.2
YOLOv10s	76.52	63.39	71.97	47.01	8.04	24.5
YOLOv9t	78.85	56.08	67.21	45.38	1.73	6.4
YOLOv9s	79.25	63.15	71.93	48.62	6.20	22.1
YOLOv8n	75.04	62.04	69.24	45.18	2.69	6.8
YOLOv8s	76.56	64.90	72.30	47.70	9.83	23.4
YOLO-FOD	74.9	57.6	66.6	37.6	6.7	16.0
YOLOv8-STE	/	/	73.51	46.41	20.3	48.3
Fog-YOLO11n	81.51	63.39	72.96	45.74	2.14	5.66
DMSM-YOLO	76.59	63.62	71.67	46.01	2.56	7.1

Note: Results of the SOTA methods are taken from the original publications. Differences exist among the SOTA methods in dataset settings, training data, input resolution, evaluation metrics, and speed measurement protocols. "/" denotes information or metrics not reported in the original publications.

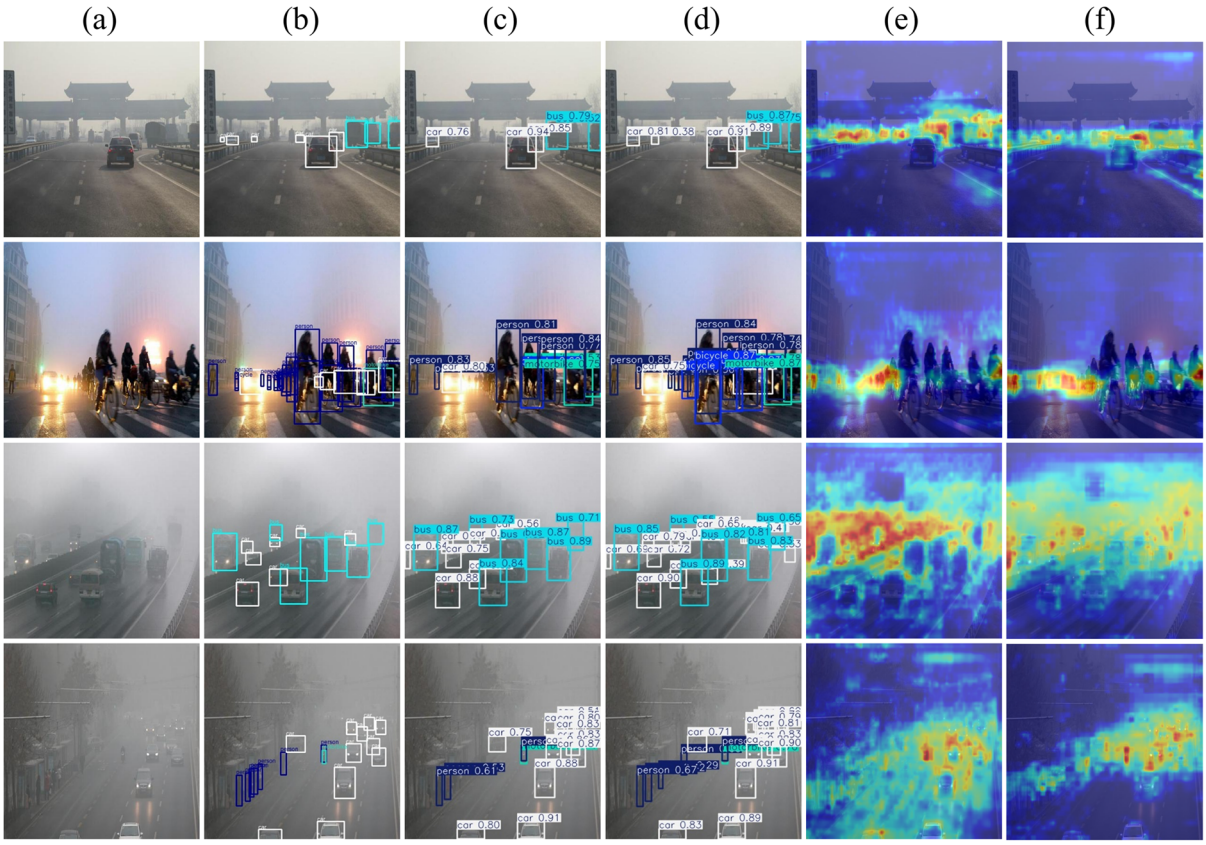


Figure 11. Detection results on the RTTS test dataset. (**a**) Original image; (**b**) ground truth; (**c**) YOLOv11n; (**d**) DMSM-YOLO; (**e**) DMSM-YOLO Grad-CAM; (**f**) YOLOv11n Grad-CAM.

3.7. Results Analysis

To evaluate the sensitivity of DMSM-YOLO to dataset partitioning, three training-validation-test ratios (8:1:1, 7:2:1, and 6:2:2) were tested on the DAWN and RTTS datasets. As shown in Figure 12, on RTTS, the corresponding mAP@0.5 values were 71.75%, 71.67%, and 71.36%, with a maximum variation of 0.39 percentage points. On DAWN, the values were 54.46%, 54.71%, and 52.96%, with a maximum variation of 1.75 percentage points. These results indicate that DMSM-YOLO maintains relatively stable performance under different dataset partition ratios, with DAWN showing slightly larger variation than RTTS, which may be partly related to its smaller dataset size.

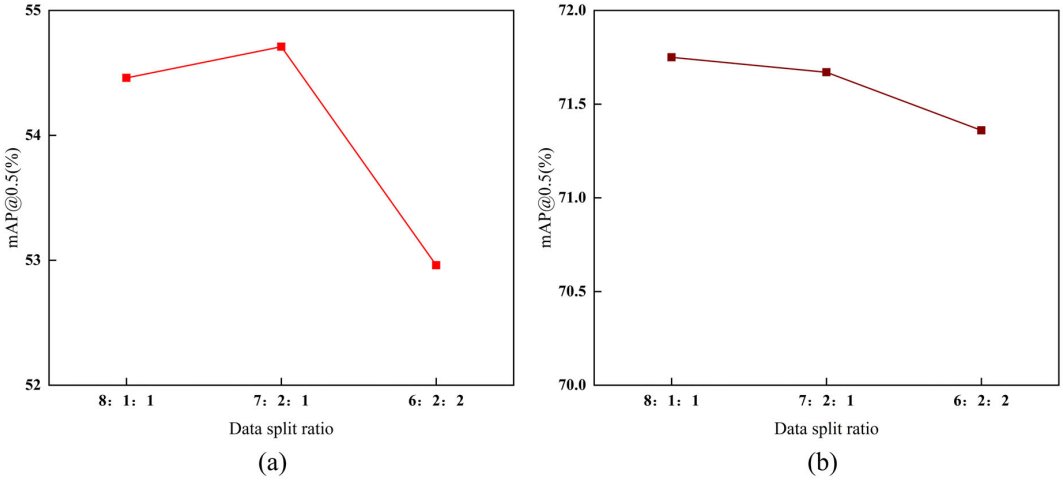


Figure 12. Sensitivity analysis of different dataset partition ratios. (**a**) DAWN; (**b**) RTTS.

As shown in Figure 13, the training loss and validation loss of DMSM-YOLO decrease rapidly in the early training stage on both DAWN and RTTS datasets. After about 100 epochs, the decline rate of loss slows down obviously, and the validation loss gradually stabilizes without an obvious upward rebound phenomenon, which indicates that the model does not suffer from severe over-fitting. Meanwhile, the mAP@0.5 curve keeps rising in the early phase and gradually converges to a saturated high-value platform in the later training period. Although slight fluctuations exist due to batch-wise random sampling, the overall detection performance tends to be stable. The above results demonstrate that DMSM-YOLO achieves favorable convergence performance on both datasets within 200 training epochs.

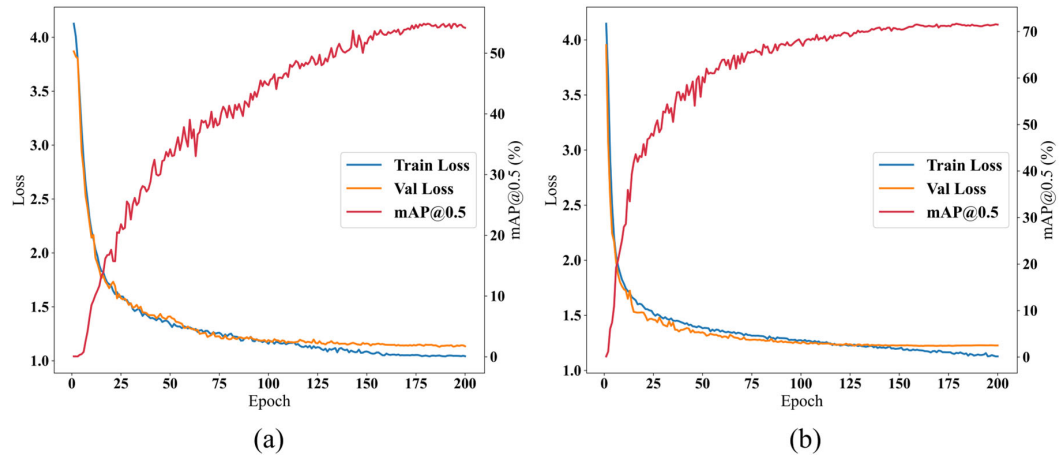


Figure 13. Training and validation curves of DMSM-YOLO. (**a**) DAWN; (**b**) RTTS.

Figure 14a presents the P-R curves of DMSM-YOLO on the DAWN dataset, where the overall mAP@0.5 reaches 0.5471. The car class achieves the highest AP of 0.8679, followed by truck with 0.5536, while motorcycle + bicycle, bus, and person obtain relatively lower AP values. This is mainly because small targets and pedestrians are more easily affected by blurred contours, occlusion, and background confusion under adverse weather conditions. Figure 14b shows the P-R curves on the RTTS dataset, where the overall mAP@0.5 reaches 0.7167. The car and person classes achieve AP values of 0.8426 and 0.7741, respectively, while bicycle, motorbike, and bus obtain AP values of 0.6686, 0.6652, and 0.6329. These results indicate that DMSM-YOLO maintains effective detection performance on both mixed adverse-weather and foggy low-visibility datasets.

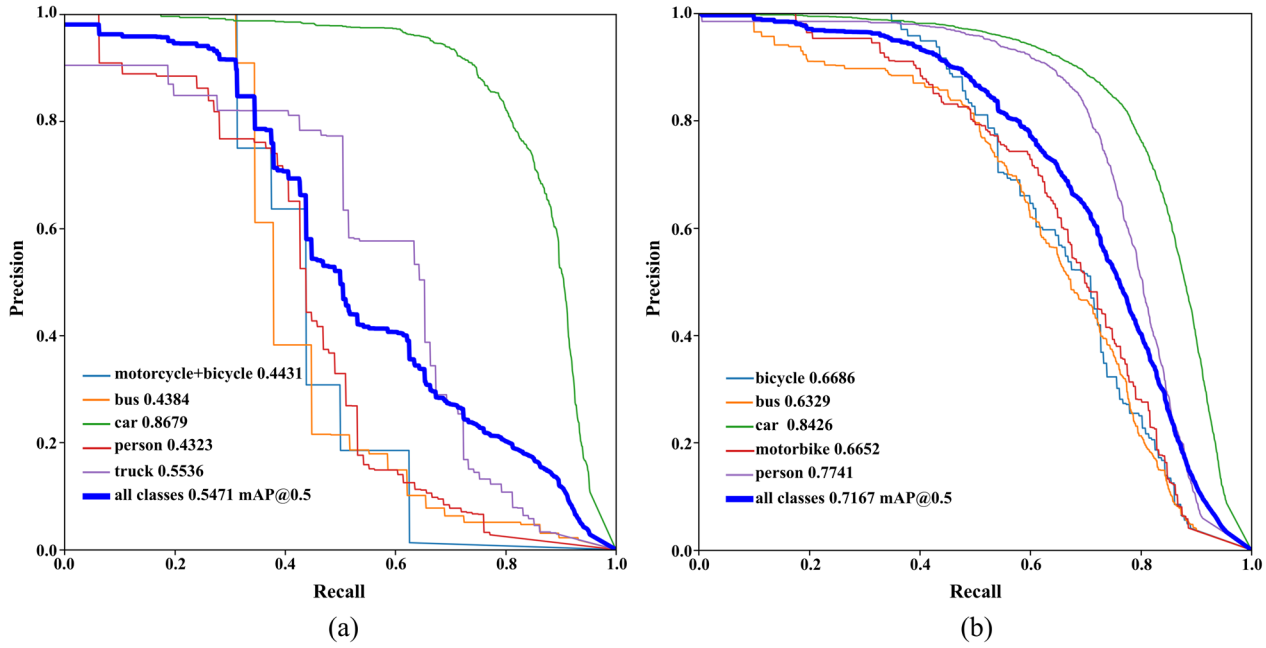
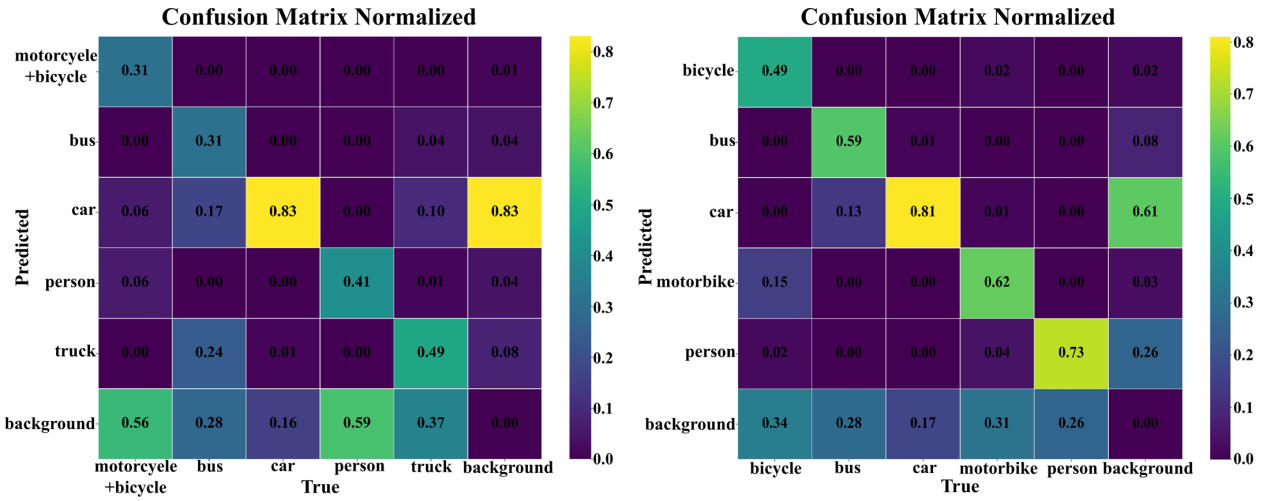


Figure 14. Precision–Recall curves on the DAWN and RTTS datasets. (**a**) DAWN; (**b**) RTTS.

Figure 15 shows the normalized confusion matrices of DMSM-YOLO on the DAWN and RTTS datasets. In Figure 15a, the car category on DAWN achieves the highest recall of 0.83, indicating that the model can effectively recognize large and structurally clear targets under mixed adverse-weather conditions. However, truck, person, bus, and motorcycle + bicycle obtain lower recall values of 0.49, 0.41, 0.31, and 0.31, respectively, and many small or blurred targets are misclassified as background. This indicates that feature degradation and background confusion remain challenging. In Figure 15b, the car and person categories on RTTS achieve recall values of 0.81 and 0.73, respectively, showing that DMSM-YOLO maintains effective recognition for major categories in foggy scenes. However, bicycle has a lower recall of 0.49 and is partially confused with motorbike, which may be caused by their similar appearance and weakened object boundaries under foggy conditions. Overall, the confusion matrices indicate that DMSM-YOLO performs effectively for major categories under low-visibility conditions, but small targets and visually similar categories remain difficult. Improving feature discrimination for weak and similar targets will be an important direction for future work.



(a) (b)

Figure 15. Confusion matrices on the DAWN and RTTS datasets. (**a**) DAWN; (**b**) RTTS.

To evaluate the statistical reliability and robustness of the proposed method against random initialization, we conducted three independent training runs with different random seeds (42, 123, 2026) on both the DAWN and RTTS datasets. Table 8 presents the random seed robustness analysis across three independent runs. On both datasets, DMSM-YOLO consistently outperforms YOLOv11n, achieving mean mAP@0.5 of 54.57% on DAWN and 71.80% on RTTS, compared to 49.06% and 68.78% for the baseline. The standard deviations across all runs remain relatively small, with the largest being 0.78% on DAWN, indicating stable training behavior regardless of random seed selection. Paired *t*-test *p*-values of 0.0103 (DAWN) and 0.0067 (RTTS) further confirm that the improvements are statistically significant. These results demonstrate that DMSM-YOLO maintains robust and consistent performance under different random initializations.

Table 8. Random seed robustness analysis on the DAWN and RTTS datasets.

	DAWN			**RTTS**		
Seed	42	123	2026	42	123	2026
YOLOv11n (mAP@0.5/%)	48.92	49.28	48.97	68.60	68.95	68.79
DMSM-YOLO (mAP@0.5/%)	55.12	53.68	54.92	71.60	72.41	71.39
Mean ± Std	49.06 ± 0.20		54.57 ± 0.78	68.78 ± 0.18		71.80 ± 0.54
p-value		0.0103			0.0067	

To evaluate the statistical uncertainty of the detection performance, Bootstrap resampling was performed on the fixed test sets of DAWN and RTTS with 1000 iterations. As shown in Table 9, the 95% confidence intervals of mAP@0.5 are [52.03%, 60.32%] on DAWN and [68.12%, 73.93%] on RTTS, respectively. These confidence intervals further characterize the variability of the reported mAP@0.5 under test-set resampling, providing additional statistical evidence for the reliability of DMSM-YOLO's detection performance.

Table 9. Statistical tests for DMSM-YOLO performance via Bootstrap resampling.

Statistical Test	Metric	Data	Result (mAP@0.5/%)
Bootstrap Test	95% CIs of mAP@0.5	DAWN RTTS	[52.03%, 60.32%] [68.12%, 73.93%]

3.8. Additional Experiments on Low-Light Scenes

To further verify the effectiveness of the proposed improvements under nighttime low-light conditions, 2, 012 nighttime low-light images from BDD100K [37] were selected for additional validation. This subset contains eight categories, including car, pedestrian, traffic light, traffic sign, bicycle, bus, truck, and rider, and was divided into training, validation, and test sets at a ratio of 7:2:1. As shown in Table 10, DMSM-YOLO achieves 49.02% Precision, 33.20% Recall, 32.11% mAP@0.5, and 15.04% mAP@0.5:0.95. Compared with YOLOv11n, Precision, Recall, mAP@0.5, and mAP@0.5:0.95 are improved by 0.30%, 2.19%, 3.12%, and 1.74%, respectively. These results indicate that the proposed improvements remain effective under nighttime low-light conditions.

Table 10. Additional validation results on the BDD100K low-light subset.

Model	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%
YOLOv11n	48.72	31.01	28.99	13.30
DMSM-YOLO	49.02	33.20	32.11	15.04

Figure 16 compares the detection results of YOLOv11n and DMSM-YOLO on the BDD100K low-light subset. Figure 16a shows the original low-light test images, and Figure 16b presents the corresponding ground-truth annotations. As shown in Figure 16c, in nighttime urban scenes, YOLOv11n misses some distant and low-contrast targets, while DMSM-YOLO detects more vehicles, pedestrians, and traffic signs with higher confidence, as shown in Figure 16d. Compared with the baseline YOLOv11n in Figure 16f, the improved DMSM-YOLO in Figure 16e produces stronger responses in object regions and less activation in irrelevant background areas, demonstrating better feature representation and target-focused capability under low-light conditions. These visual results further confirm the effectiveness of the proposed improvements under low-light conditions.

3.9. Discussion

The experimental results show that the improvements introduced in DMSM-YOLO operate at different stages of the network and contribute positively to object detection under low-visibility conditions. EPConv enhances directional contour interaction, C3k2-MDFI strengthens multi-scale weak-feature representation, while MLCA and SMDetect further refine fused features and prediction features. The performance improvements achieved on both DAWN and RTTS, together with the additional gains obtained on the BDD100K low-light subset, indicate that DMSM-YOLO remains effective under different types of low-visibility degradation. These improvements are essentially designed to preserve target-related information while suppressing interference caused by visual degradation. From this perspective, related hyperspectral target and anomaly detection studies also improve robustness by suppressing background interference while preserving target-related structural information [38,39]. Although these methods differ from RGB object detection in terms of data modality and task formulation, they share a similar feature-enhancement principle with DMSM-YOLO. Unlike explicit low-rank, sparse, or dictionary-based representation strategies, DMSM-YOLO directly enhances degraded visual features through directional

contour interaction in EPConv and multi-scale feature modeling in MDFI, integrating feature enhancement into a lightweight object detection network. From the perspective of feature representation, data augmentation introduces visual diversity while preserving target-related semantic information. Therefore, the intrinsic relationship between original and augmented samples can be understood in terms of representation consistency under appearance perturbations. Previous work [40] has examined the similarity between original and augmented representations, providing a useful perspective for analyzing this relationship. For low-visibility object detection, appropriate data augmentation may encourage the model to learn features that are less sensitive to variations in illumination, contrast, and appearance, thereby contributing to improved generalization, robustness, and training stability.

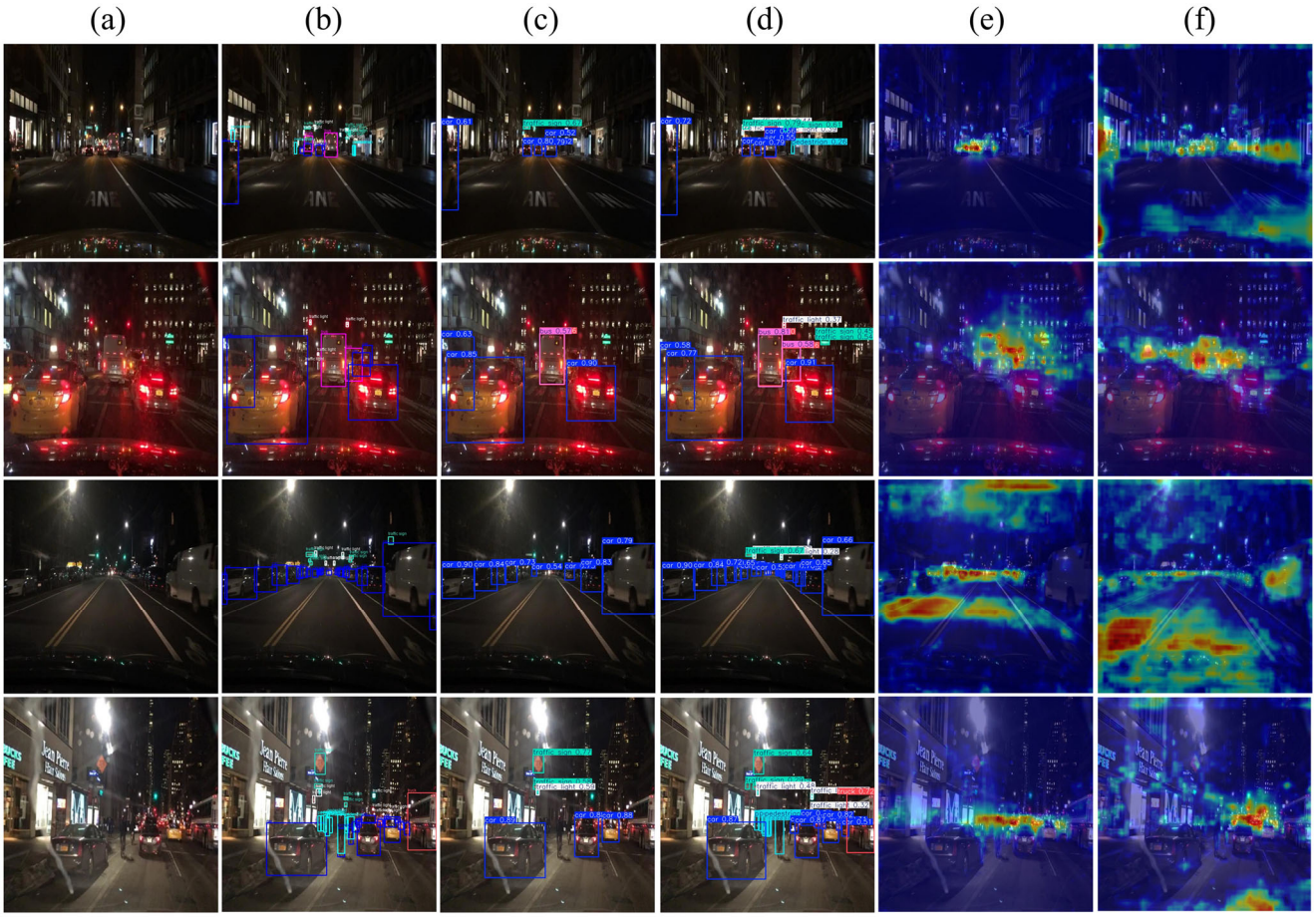


Figure 16. Detection results on the BDD100K low-light test dataset. (**a**) Original image; (**b**) ground truth; (**c**) YOLOv11n; (**d**) DMSM-YOLO; (**e**) DMSM-YOLO Grad-CAM; (**f**) YOLOv11n Grad-CAM.

Despite these improvements, the current study still has certain limitations. To begin with, DAWN, RTTS, and the BDD100K low-light subset are mainly derived from general road scenes. The visual degradation characteristics contained in these datasets, including reduced contrast, blurred object boundaries, and weakened target features, are similar to those encountered in outdoor AGV perception under low-visibility conditions, but they cannot fully represent real outdoor AGV environments. Therefore, the current results mainly demonstrate the robustness of the model to low-visibility degradation and should not be regarded as complete evidence of cross-domain generalization to industrial AGV scenarios. Furthermore, because some SOTA results are taken from the original publications and obtained under different experimental protocols, the corresponding comparisons

should be interpreted as contextual references rather than strictly controlled head-to-head evaluations. In addition, the reported inference speed was measured on an RTX 4090D workstation and cannot directly represent deployment performance on embedded AGV platforms. Future work will focus on collecting representative industrial AGV data, conducting cross-domain evaluations, and validating the model on embedded platforms.

4. Conclusions

To improve detection performance for outdoor AGV visual perception under low-visibility conditions, DMSM-YOLO is proposed as an improved lightweight object detection model based on YOLOv11n. First, EPConv is designed with diagonal residual interaction to enhance directional feature collaboration and improve blurred-boundary representation. Second, MDFI is designed to construct C3k2-MDFI, thereby strengthening weak-target feature extraction and improving the representation of blurred and low-contrast objects. MLCA is introduced after feature-fusion operations to recalibrate low-contrast fused features and suppress background interference. Finally, SimAM is introduced into the detection head to construct SMDetect, improving localization accuracy and confidence estimation for blurred and occluded objects. Experimental results show that DMSM-YOLO achieves mAP@0.5 values of 54.71% and 71.67% on the DAWN and RTTS datasets, respectively, which are 5.36 and 3.26 percentage points higher than those of YOLOv11n. Meanwhile, the model maintains 2.56 M parameters and 7.1 GFLOPs. Additional low-light experiments further verify the effectiveness of the proposed method under nighttime low-visibility conditions. Overall, DMSM-YOLO improves detection accuracy and robustness while maintaining low model complexity, providing a lightweight detection solution for AGV visual perception under low-visibility conditions.

Author Contributions: Funding acquisition, T.W.; Writing—original draft, H.Z.; Conceptualization, S.C. (Shuwan Cui); Methodology, G.Z.; Writing—review and editing, S.C. (Shibing Cai); Software, Y.L.; Validation, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Guangxi Key Technologies R&D Program (Grant No. AB24010207).

Data Availability Statement: Data will be provided upon reasonable request. The data are not publicly available due to their inclusion in the article.

Conflicts of Interest: The authors declare no conflicts of interest. Authors Tao Wei, Yilong Li and Shaodong Zheng were employed by the company Guangxi Nanguo Copper Industry Co., Ltd.; Author Huanwu Zhan was employed by the company Nandan Nanfang Nonferrous Metals Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bhargava, A.; Suhaib, M.; Singholi, A.S. A review of recent advances, techniques, and control algorithms for automated guided vehicle systems. *J. Braz. Soc. Mech. Sci. Eng.* **2024**, *46*, 419. [CrossRef]
2. Song, Y.; Lu, Y. A Review of Unmanned Visual Target Detection in Adverse Weather. *Electronics* **2025**, *14*, 2582. [CrossRef]
3. Chen, Z.; Zhang, Z.; Su, Q.; Yang, K.; Wu, Y.; He, L.; Tang, X. Object Detection for Autonomous Vehicles Under Adverse Weather Conditions. *Expert Syst. Appl.* **2026**, *296*, 128994. [CrossRef]
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]

6. Wang, W.; Jing, B.; Yu, X.; Sun, Y.; Yang, L.; Wang, C. YOLO-OD: Obstacle Detection for Visually Impaired Navigation Assistance. *Sensors* **2024**, *24*, 7621. [CrossRef]
7. Cui, S.; Liu, F.; Wang, Z.; Zhou, X.; Yang, B.; Li, H.; Yang, J. DAN-YOLO: A Lightweight and Accurate Object Detector Using Dilated Aggregation Network for Autonomous Driving. *Electronics* **2024**, *13*, 3410. [CrossRef]
8. Cai, L.; Wang, H.; Zhou, C.; Wang, Y.; Liu, B. MFMAM-YOLO: A Method for Detecting Pole-Like Obstacles in Complex Environments. *Neural Comput. Appl.* **2025**, *37*, 12673–12698. [CrossRef]
9. Sun, Y.; Lu, X.; Lei, L. Improved YOLOv5 Model for Pedestrian and Vehicle Detection. In *Proceedings of the 2023 3rd International Conference on Intelligent Communications and Computing (ICC)*; IEEE: New York, NY, USA, 2023; pp. 1–5. [CrossRef]
10. Yang, Q.; Lian, Y.; Liu, Y.; Xie, W.; Yang, Y. Multi-AGV Tracking System Based on Global Vision and AprilTag in Smart Warehouse. *J. Intell. Robot. Syst.* **2022**, *104*, 42. [CrossRef]
11. Yang, D.; Su, C.; Wu, H.; Xu, X.; Zhao, X. Shelter Identification for Shelter-Transporting AGV Based on Improved Target Detection Model YOLOv5. *IEEE Access* **2022**, *10*, 119132–119139. [CrossRef]
12. Zhang, K.; Yang, M. Research on Improved YOLOv5s Algorithm Based on Optical Sensors for AGV Object Detection. In *Proceedings of the Second International Conference on Intelligent Transportation and Smart Cities (ICITSC 2025)*; SPIE: Bellingham, WA, USA, 2025; Volume 13682, p. 46. [CrossRef]
13. Tan, Z.; Chu, W.; Zhu, J.; Zhu, G.; Li, H.; Hua, S.; Li, B.; Zhang, F.; Teng, S.; Pan, S. Obstacle Avoidance Technology of AGV for Power Construction Based on an Improved YOLOv8 Algorithm. *J. Phys. Conf. Ser.* **2026**, *3215*, 012020. [CrossRef]
14. Cai, S.; Wu, Z.; Liu, K.; Zhang, T.; Weng, W.; Zheng, X. LSOD-YOLO: A Visual Object Detection Method for AGV Perception Systems Based on a Lightweight Backbone and Detection Head. *Technologies* **2026**, *14*, 173. [CrossRef]
15. Tai, J.; Zhao, J.; Chen, D.; Shi, Y. AGV Dynamic Recognition Algorithm based on Optimized YOLOv3. In *Proceedings of the 2023 IEEE 11th International Conference on Information, Communication and Networks (ICICN)*; IEEE: New York, NY, USA, 2023; pp. 708–713.
16. Wu, T.; Zhang, Y.; Zhao, H.; Yue, Y.; Yu, L.; Wang, X. Enhancing Automated Guided Vehicle Navigation with Multi-Sensor Fusion and Algorithmic Optimization. In *Proceedings of the 2024 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM)*; IEEE: New York, NY, USA, 2024; pp. 557–562. [CrossRef]
17. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
18. Ultralytics. Ultralytics YOLOv8. GitHub Repository, 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 1 June 2026).
19. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
20. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458.
21. Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; Huang, X. Pinwheel-shaped Convolution and Scale-based Dynamic Loss for Infrared Small Target Detection. *arXiv* **2024**, arXiv:2412.16986.
22. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the Stars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 5694–5703. [CrossRef]
23. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed Local Channel Attention for Object Detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [CrossRef]
24. Yang, L.; Zhang, R.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; pp. 11863–11874.
25. Kenk, M.; Hassaballah, M. DAWN: Vehicle detection in adverse weather conditions dataset. *arXiv* **2020**, arXiv:2008.05402.
26. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking Single-Image Dehazing and Beyond. *IEEE Trans. Image Process.* **2019**, *28*, 492–505. [CrossRef]
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542. [CrossRef]
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722. [CrossRef]
30. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
31. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. *arXiv* **2025**, arXiv:2509.25164.

32. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
33. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
34. Liu, Y.; Yuan, T.; Ren, A.; Kuo, Y.; Xiong, X. YOLO-FOD: Lightweight object detection based on multibranch and multiscale feature fusion for adverse weather. *Neurocomputing* **2026**, *659*, 131778. [CrossRef]
35. Jing, Z.; Li, S.; Zhang, Q. YOLOv8-STE: Enhancing Object Detection Performance Under Adverse Weather Conditions with Deep Learning. *Electronics* **2024**, *13*, 5049. [CrossRef]
36. Kuang, F.; Chen, Z. Fog-YOLO11n: A Lightweight Traffic Object Detection Framework for Autonomous Driving Under Foggy Conditions. *Appl. Sci.* **2026**, *16*, 8059. [CrossRef]
37. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645. [CrossRef]
38. Lin, S.; Cai, E.; Zeng, Y.; Zhu, Y.; Zhao, W.; Zhang, L.; Cao, B. A hyperspectral target detection method by combining robust dictionaries and constrained energy minimization regularized low-rank and sparse representation model. *Adv. Space Res.* **2026**, *78*, 3523–3544. [CrossRef]
39. Lin, S.; Zhang, M.; Cheng, X.; Shi, L.; Gamba, P.; Wang, H. Dynamic Low-Rank and Sparse Priors Constrained Deep Autoencoders for Hyperspectral Anomaly Detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2500518. [CrossRef]
40. Zeng, Y.; Wang, X.; Cheng, Y. Similarity-guided state attention for visual reinforcement learning. *Neural Netw.* **2026**, *205*, 109560. [CrossRef]