YOLOv11-TWCS: Enhancing Object Detection for Autonomous Vehicles in Adverse Weather Conditions Using YOLOv11 with Self-Attention

Chris Michael and Hongjian Wang *

Posted Date: 25 November 2025

doi: 10.20944/preprints202511.1864.v1

Keywords: attention mechanism; deep learning; YOLOv11; adverse weather; object detection; trans weather; CBAM; SCDown

 doi:10.20944/preprints202511.1864.v1

Article

YOLOv11-TWCS: Enhancing Object Detection for Autonomous Vehicles in Adverse Weather Conditions Using YOLOv11 with Self-Attention

Chris Michael [1], and Hongjian Wang [1,*]

School of Computer Science and Technology, Donghua University, Shanghai 201620, China
* Correspondence: hongjian.wang@dhu.edu.cn

Abstract

Object detection for autonomous vehicles under adverse weather conditions—such as rain, fog, snow, and low light—remains a significant challenge due to severe visual distortions that degrade image quality and obscure critical features. This paper presents YOLOv11-TWCS, an enhanced object detection model that integrates TransWeather, the Convolutional Block Attention Module (CBAM), and Spatial-Channel Decoupled Downsampling (SCDown) to improve feature extraction and emphasize critical features in weather-degraded scenes while maintaining real-time performance. Our approach addresses the dual challenges of weather-induced feature degradation and computational efficiency by combining adaptive attention mechanisms with optimized network architecture. Evaluations on DAWN, KITTI, and Udacity datasets show improved accuracy over baseline YOLOv11 and competitive performance against other state-of-the-art methods, achieving mAP@0.5 of 59.1%, 81.9%, and 88.5%, respectively. The model reduces parameters and GFLOPs by approximately 19–21% while sustaining high inference speed (105 FPS), making it suitable for real-time autonomous driving in challenging weather conditions.

Keywords: attention mechanism; deep learning; YOLOv11; adverse weather; object detection; trans weather; CBAM; SCDown

1. Introduction

In recent years, autonomous driving technology has advanced rapidly, making computer vision a key component for self-driving systems [1]. However, detecting vehicles accurately in complex environments remains a significant challenge. Traditional methods relied on manual feature extraction, which is time-consuming and often less robust. Deep learning-based object detection algorithms have greatly improved performance by automatically learning features from data [2]. These algorithms typically perform well under favorable weather and lighting conditions, while vehicle detection under adverse conditions, such as rain, fog, and low light, is still difficult. This paper aims to address this challenge by improving detection accuracy in such complex scenarios while maintaining real-time performance.

Deep learning-based object detection algorithms are generally divided into two types: two-stage detectors, which first generate candidate regions and then classify them, and one-stage detectors, which predict object locations and classes in a single step. Despite these advances, in the two-stage detection framework, candidate bounding boxes are first generated, followed by refinement for final detection. Representative models include R-CNN [3], Faster R-CNN [4], and Mask R-CNN [5]. In contrast, one-stage detection algorithms, such as SSD [6,7] and the YOLO (You Only Look Once) family [8,9], predict bounding boxes and class probabilities in a single pass. Despite advances, these methods often impose a heavy computational load, especially on edge devices with limited resources [10]. The emergence of YOLO, with its real-time capabilities and simplified architecture, significantly improved detection

 doi:10.20944/preprints202511.1864.v1

efficiency by eliminating the region proposal stage [9]. Further, vehicle detection remains challenging, particularly under complex scenes and harsh weather. Visual issues due to varying viewpoints, occlusions, and truncations are common [11]. Adverse conditions such as rain, fog, snow, and low light further hinder the detection performance by introducing noise and reducing image quality. With the rapid progress in deep learning, the YOLO family of models [9] has become a mainstream choice for high-speed object detection. The most recent iteration, YOLOv11 [12], incorporates architectural enhancements and training improvements to push performance boundaries. However, road object detection still suffers in adverse weather, with rain notably degrading image clarity and object visibility.

Considering the challenges posed by adverse weather conditions and aiming to enhance detection capabilities in real-world scenarios, we propose YOLOv11-TWCS, a novel detection framework built upon the YOLOv11m architecture. It is designed to perform robustly under conditions such as rain, fog, snow, and low light while maintaining real-time performance. This work enhances the YOLOv11 architecture to improve feature extraction on noisy and small-scale datasets, which are commonly encountered in adverse weather conditions, without compromising computational efficiency. YOLOv11-TWCS effectively addresses the limitations of current object detectors in challenging environments. The main contributions of this study are summarized as follows:

- To improve feature extraction from objects that are blurred or occluded by rain, snow, fog, or sandstorms, we integrated the TransWeather Block [13,14] into the PSA block of C2PSA in the final layer of the backbone network in YOLOv11, enabling the model to capture both local and global features effectively.
- Within the Neck network, we incorporated the Convolutional Block Attention Module (CBAM) [15] following each C3k2 layer to highlight salient features in the feature maps while diminishing less relevant information via channel-wise and spatial attention mechanisms.
- To reduce the computational cost introduced by TransWeather and CBAM without sacrificing detection accuracy, we substituted selected convolutional layers in the backbone network with Spatial-Channel Decoupled Downsampling (SCDown) modules [16,17]. These modules make the model more efficient by compactly reducing both the spatial and channel dimensions of the feature maps, while simultaneously enhancing detection performance.
- The proposed YOLOv11-TWCS model was extensively evaluated on the DAWN dataset using key metrics such as mAP@0.5, FPS, GFLOPs, and parameter count. Experimental results and ablation studies verify that each integrated module significantly enhances detection accuracy and computational efficiency. Its strong generalization capability is further validated through reliable results on the Udacity and KITTI datasets under both clear and low-light environments.

The rest of this paper is organized as follows. Section 2 introduces related work on object detection in adverse conditions. Section 3 describes our methodology including the YOLOv11 model and our proposed YOLOv11-TWCS architecture. Section 4 presents experiments and analysis, and Section 5 concludes our paper.

2. Related Work

Most models for object detection under challenging lighting and adverse weather conditions have attracted considerable attention due to their critical role in real-world applications such as autonomous driving and surveillance. Early object detection in intelligent transportation was dominated by traditional, hand-crafted approaches that relied on engineered features and classical classifiers. These methods extracted local interest points or corner features and then used rule-based or machine learning classifiers to decide whether a region contained a vehicle; however, they suffered from low accuracy, poor robustness in fuzzy scenes, and high computational cost.

The shift to deep learning began with two-stage detectors that prioritized accuracy by generating region proposals before classification. Ren et al. [4] proposed Faster R-CNN, which combined region proposal networks with convolutional backbones to greatly improve detection precision across scales. Although highly accurate, two-stage pipelines introduced latency that limited real-time use in driving.

 doi:10.20944/preprints202511.1864.v1

Single-stage detectors then emerged to trade a small loss in peak accuracy for much higher speed. Liu et al. introduced SSD [6,7], a single-shot multibox detector that enabled faster inference through multi-scale prediction on convolutional feature maps. Redmon's YOLO family further prioritized real-time performance; YOLOv4 (Bochkovskiy et al. [18]) combined CSPDarknet, path aggregation, and optimized activations (Mish) to deliver a strong speed–accuracy balance for automotive vision.

To improve feature representation and robustness without sacrificing speed, researchers proposed architectural modules and fusion strategies. Liu et al. [19] introduced RFBNet, adding Receptive Field Blocks inspired by human visual eccentricity to SSD, thereby strengthening multi-scale receptive fields and improving discriminability for small and varied objects while keeping real-time throughput. TPH-YOLOv5 (Zhu et al. [20]) augmented the YOLO family with a transformer-based prediction head to capture broader context and global relationships—beneficial for complex scenes such as aerial/drone views.

As transformer-based models matured, they were adapted for vision tasks to capture long-range dependencies. Liu et al. [21] proposed the Swin Transformer, which introduced hierarchical shifted-window self-attention to model both local and global context efficiently. Building on attention paradigms, RT-DETR variants (Zhao et al. [22,23]) integrated dynamic label assignment and efficient transformer encoders to improve small-object detection while maintaining real-time operation.

Concurrently, architecture and training refinements targeted both efficiency and robustness. Lyu et al. [24] developed RTMDet, which uses large-kernel depth-wise convolutions and better optimization strategies (including soft labels in dynamic assignment) to push throughput beyond prior YOLO variants while sustaining accuracy. Loss and label strategies, plus kernel design, helped close the gap between speed and high performance.

Recent application-driven studies enhance YOLO architectures for adverse and low-visibility scenes. Liu et al.[25] developed SPR-YOLO, a lightweight YOLOv8/YOLOv8-MobileNetV3 variant with SPD-Conv, SECA attention, and DY-GELAN fusion for improved feature extraction in noisy or rainy conditions. Tahir et al. [26] proposed PVDM-YOLOv8l, integrating the Swin Transformer, CBAM, and Wise-IoU v3 loss, achieving higher detection accuracy under poor visibility using the enhanced DAWN2024 dataset. Similarly, Sun et al. combined dehazing preprocessing with Swin-based detectors, and Tahir et al. (in Drones) introduced PVswin-YOLOv8s, leveraging transformer and attention modules for robust UAV-based pedestrian and vehicle detection.

While these approaches have significantly advanced object detection under adverse conditions through innovations in feature enhancement, attention mechanisms, preprocessing techniques, and specialized loss functions, they still face persistent challenges in maintaining an optimal balance between precision and recall. This is particularly evident when detecting small, distant, or partially occluded objects where visible features are inherently limited. Such limitations underscore the critical need for more integrated frameworks that not only improve overall detection accuracy but also specifically enhance capabilities for identifying subtle, obscured, and low-confidence targets—a research gap that our work directly addresses to achieve more reliable and robust performance in complex real-world environments.

3. Methodology

3.1. Basic YOLOv11 Model

YOLOv11, developed by the Ultralytics team, marks a significant evolution in the YOLO series by introducing architectural innovations that focus on improving both accuracy and efficiency. The model replaces the conventional C2f module with the C3k2 module, which enhances gradient propagation and deep feature extraction. Additionally, it incorporates the Channel-to-Pixel Space Attention (C2PSA) module and Spatial Pyramid Pooling Fusion (SPPF), allowing the network to better capture multiscale context and spatial relationships, making it well-suited for complex visual environments. The improvements in YOLOv11 enable it to achieve a 22% reduction in parameters over YOLOv8 and

 doi:10.20944/preprints202511.1864.v1

a 2% speed increase over YOLOv10. It also extends its capabilities to support multitask processing such as pose estimation and instance segmentation, making it versatile for real-world applications.

Despite these advances, YOLOv11 faces challenges when operating in visually degraded scenes, particularly those affected by adverse weather conditions such as rain, fog, snow, and low illumination. These environments often result in feature ambiguity, contrast degradation, and motion blur factors that diminish detection accuracy in real-world deployments like autonomous driving.

3.2. Overview of YOLOv11-TWCS Model

The YOLOv11-TWCS builds upon the YOLOv11 foundation by enhancing its capacity to perceive and process degraded visual inputs. In the final part of the backbone network, we introduce the TransWeather Block, into the PSA component of the C2PSA module. This block integrates Dynamic-Range Self-Attention (DRSA), which captures subtle intensity variations influenced by environmental noise, and Dual-Scale Gated Feed-Forward (DGFF), which refines contextual information across multiple resolutions. In the Neck network, YOLOv11-TWCS integrates the Convolutional Block Attention Module (CBAM) after each C3k2 layer to highlight salient features while suppressing less relevant information through channel and spatial attention. To balance the additional computational cost introduced by TransWeather and CBAM, selected convolutional layers in the backbone are replaced with Spatial-Channel Decoupled Downsampling (SCDown) modules, which compactly reduce both spatial and channel dimensions. This design enhances computational efficiency while preserving high detection accuracy, as illustrated in Figure 1.

 doi:10.20944/preprints202511.1864.v1

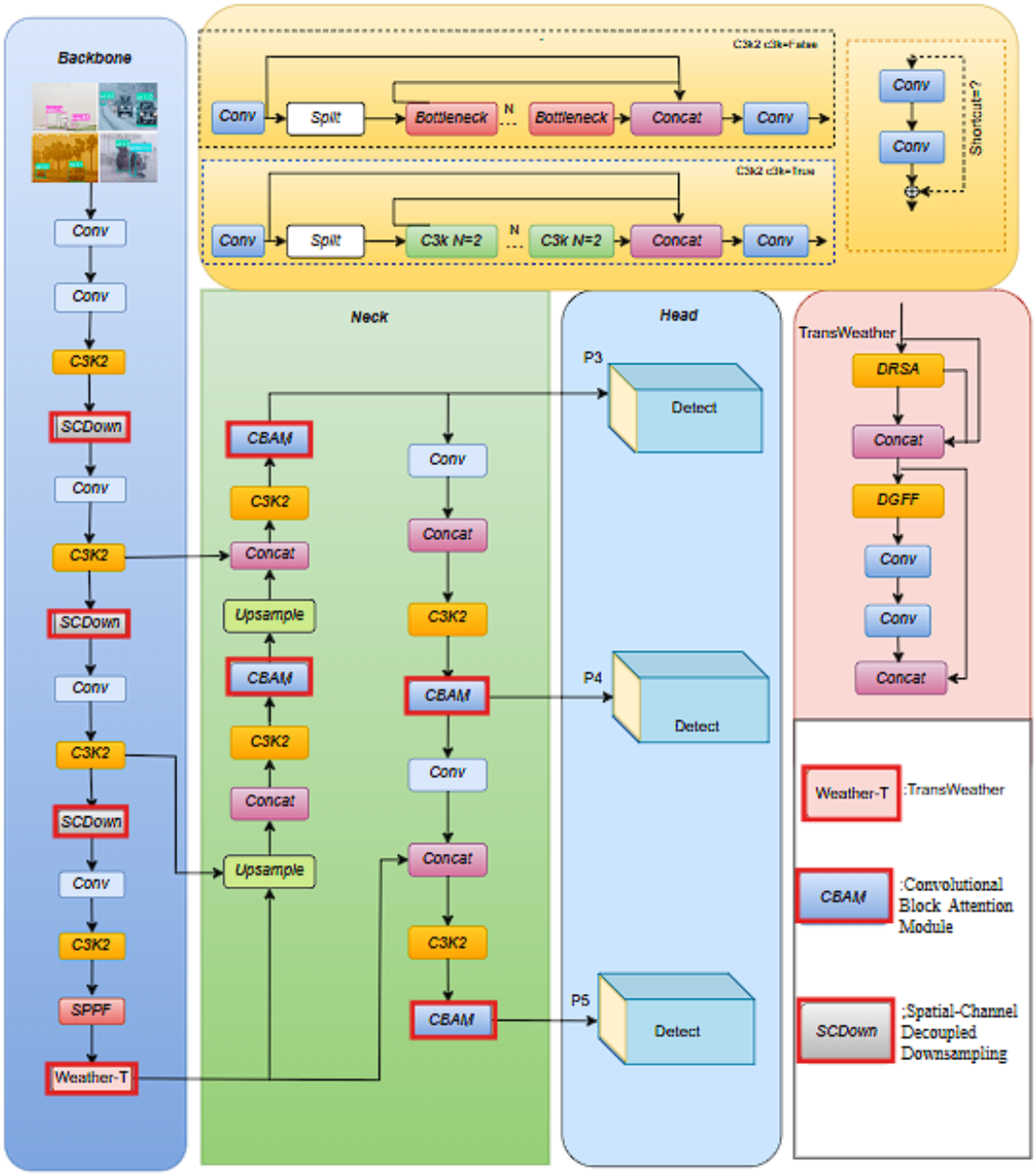

Figure 1. The network structure of YOLOv11-TWCS

These enhancements allow YOLOv11-TWCS to extract more resilient and meaningful features from distorted inputs, improving its performance in scenes characterized by poor visibility or extreme weather effects. By incorporating this adaptive attention mechanism and multiscale refinement, YOLOv11-TWCS maintains high detection accuracy while remaining lightweight and fast, meeting the demands of real-time systems in unpredictable and challenging conditions.

3.3. TransWeather Block

In our proposed YOLOv11-TWCS model, we integrate the TransWeather Block [13,14] into the PSA component of the C2PSA module to improve the extraction of features under adverse weather conditions. The block, commonly used in image restoration and transformer-based detection frameworks, dynamically adjusts attention to variations in brightness, noise, and low-contrast regions caused by rain, fog, snow, and low-light environments. Standard YOLO architectures, while efficient under normal conditions, often struggle with degraded visual inputs, resulting in missed detections or misclassification of small, distant, or partially occluded objects. By incorporating the TransWeather

 doi:10.20944/preprints202511.1864.v1

Block, our model selectively emphasizes informative pixels and suppresses weather-induced noise, improving the representation of subtle and low-confidence targets. While the block is highly effective, it can introduce additional computational overhead due to its pixel-level dynamic attention mechanism, which in standalone transformer frameworks may slightly reduce inference speed. Its design primarily targets low-quality image features, which may not directly improve detection in clear conditions. However, when integrated into YOLOv11-TWCS, the block complements YOLO's efficient backbone and neck, enhancing degraded regions without significantly affecting real-time performance. Its inclusion is particularly important for autonomous vehicle perception, as it improves detection of small, distant, or partially occluded objects in rain, fog, snow, or low-light scenarios—conditions where standard YOLO models typically struggle. By leveraging YOLO's speed and the TransWeather Block's adaptive attention, YOLOv11-TWCS achieves a practical balance between robustness in adverse weather and high inference efficiency.

The TransWeather Block is composed of two tightly coupled components: 1. Dynamic-Range Self-Attention (DRSA) 2. Dual-Scale Gated Feed-Forward (DGFF).

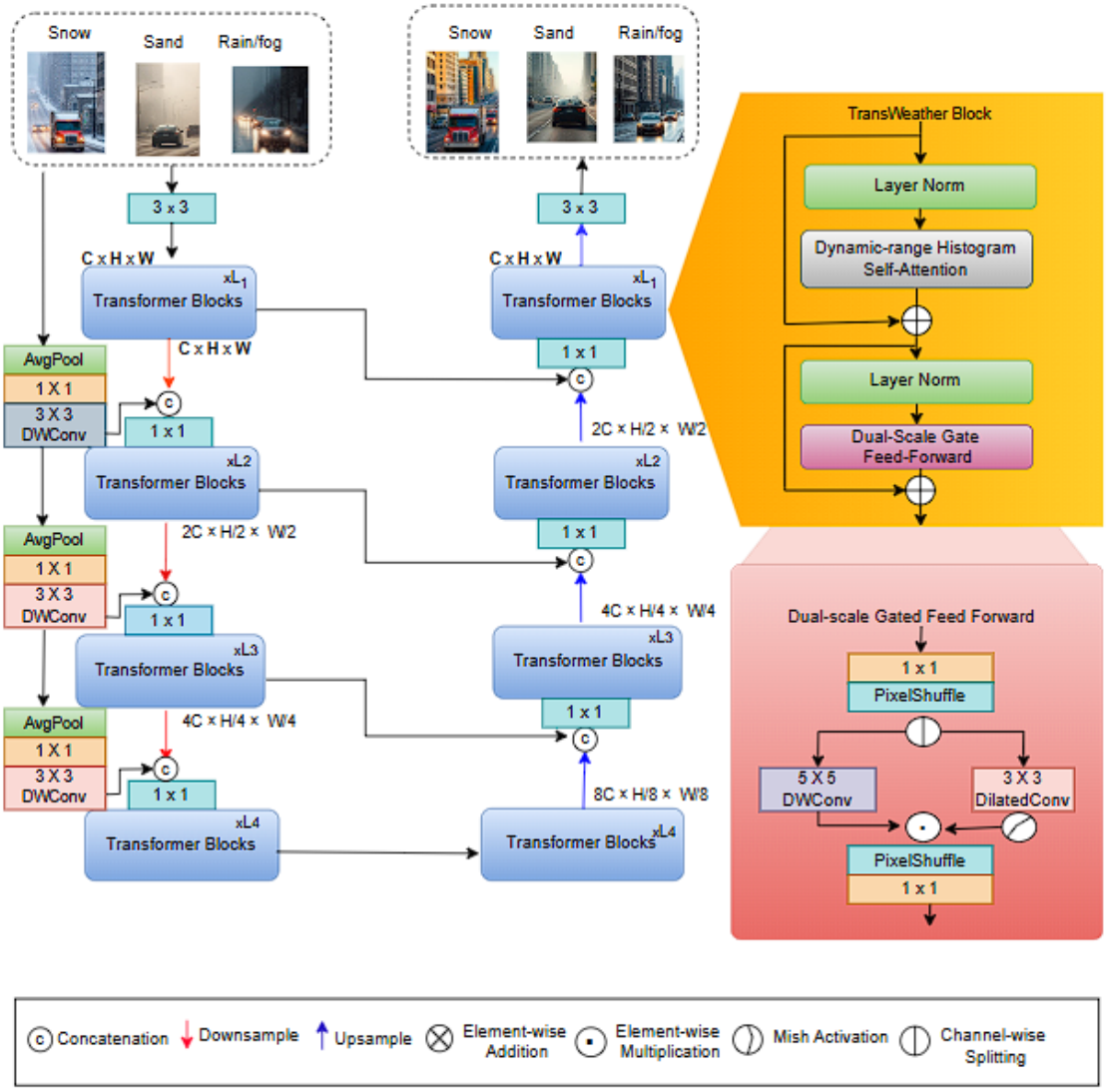

Figure 2. The structure of the TransWeather Block

The DRSA module is the first key component of the TransWeather Block, designed to enhance the network's ability to handle visually degraded inputs caused by rain, fog, snow, and low-light conditions. It works by grouping pixels into bins according to brightness levels and applying self-attention operations both within and across these bins. This structured attention mechanism enables

 doi:10.20944/preprints202511.1864.v1

the model to selectively focus on relevant features while suppressing irrelevant distortions and noise, improving the localization of small, distant, or partially occluded objects. By adapting dynamically to variations in brightness, DRSA strengthens the representation of subtle image details critical for autonomous vehicle perception in adverse weather.

The DGFF module enhances multi-scale feature representation through a gated feed-forward mechanism. It balances local detail and global context, allowing the model to retain fine object boundaries while preserving semantic consistency across the scene. This is particularly important for handling occlusions and low-contrast regions, where accurate feature aggregation is challenging. By integrating DGFF into the backbone alongside DRSA, the TransWeather Block effectively improves the resilience and accuracy of the model in complex, weather-degraded environments.

Together, these innovations significantly strengthen the model's resilience and accuracy in visually challenging scenes.Within the model's backbone, feature refinement is achieved through a two-stage process combining DRSA and the Dual-Scale Gated Feed-Forward (DGFF) module. The update mechanism for feature maps at stage 1 is mathematically described as:

$$F_l = F_{\{l-1\}} + DRSA(LN(F_{\{l-1\}})) \tag{1}$$

$$F_l = F_l + DGFF(LN(F_l)) \tag{2}$$

where LN denotes layer normalization, and F_l represents the feature map at the l-th stage. DRSA itself consists of two primary stages: dynamic-range convolution and dual-path histogram-based self-attention. In the dynamic-range convolution stage, traditional fixed-kernel convolutional operations are enhanced by introducing a sorting-based mechanism that allows for dynamic adjustment to spatial features. Given an input tensor $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, we divide it along the channel dimension into two branches F1 and F2. The first branch undergoes horizontal and vertical sorting to reorder spatial information before merging with the second branch. The merged features are then processed through a point-wise convolution followed by a depth-wise convolution:

$$F_1, F_2 = Split(F); \quad F_1 = Sort_v(Sort_h(F_1)) \tag{3}$$

$$F = Conv^d_{3\times 3}(Conv_{1\times 1}(Concat(F_1, F_2))) \tag{4}$$

where $Conv_{1\times 1}$ is point-wise convolution, $Conv^d_{3\times 3}$ is depth-wise convolution, $Concat$ is the channel-wise concatenation, $Sort_h$ and $Sort_v$ refer to horizontal and vertical sorting. This process reorganizes pixels by intensity, enabling convolutions to more effectively separate and restore degraded regions while preserving clean areas.

Following dynamic-range convolution, the output $V \in \mathbb{R}^{C\times H\times W}$ from dynamic-range convolution, and $F_{QK,1}$, $F_{QK,2} \in \mathbb{R}^{2C\times H\times W}$ represent two sets of Query-Key pairs. The features are reshaped and sorted as:

$$V, d = Sort(R^{C\times HW}_{C\times H\times W}(V)) \tag{5}$$

$$Q_1, K_1 = Split(Gather(R^{C\times HW}_{C\times H\times W}(F_{QK,1}), d)) \tag{6}$$

$$Q_2, K_2 = Split(Gather(R^{C\times HW}_{C\times H\times W}(F_{QK,2}), d)) \tag{7}$$

where $R^{C\times HW}_{C\times H\times W}$ represents the operation of reshaping features from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{C\times HW}$, d is the index of sorted value, and Gather retrieves tensor elements using index d. We apply two types of reshaping: Bin-wise Histogram Reshaping (BHR) for capturing global information and Frequency-wise Histogram Reshaping (FHR) for capturing local detail. Self-attention is then performed in each reshaping domain:

 doi:10.20944/preprints202511.1864.v1

$$A_B = softmax\left(\frac{(R_B(Q_1)R_B(K_1)^T)}{\sqrt{k}}\right)R_B(V) \tag{8}$$

$$A_F = softmax\left(\frac{(R_F(Q_2)R_F(K_2))^T}{\sqrt{k}}\right)R_B(V) \tag{9}$$

$$A = A_B \odot A_F \tag{10}$$

where R_B and R_F denote BHR and FHR reshaping operations, k is the head dimension, $\odot$ denotes element-wise multiplication. This dual-path attention structure reinforces both coarse and fine-grained contextual comprehension of weather-degraded features, enhancing the robustness of object detection. The DGFF module, applied subsequently, refines these features across multiple spatial scales using gated processing that balances local and global representations. This enables more stable detection of objects under occlusions and visual distortions, ensuring precise performance in complex real-world environments.

3.4. CBAM

The Convolutional Block Attention Module (CBAM) is composed of two sequential submodules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM) [15]. The channel attention mechanism functions by first applying global average pooling and global max pooling to the input feature map. These pooled features are then passed through a shared multi-layer perceptron (MLP), and the resulting outputs are summed before being activated using a sigmoid function. This produces a channel attention map that assigns an importance weight to each channel of the feature map. The obtained weights are then multiplied channel-wise with the input feature map to strengthen the most informative features. Attention mechanisms such as CBAM have been extensively explored to enhance object detection models by allowing them to focus selectively on meaningful regions within complex and noisy scenes [27]. To emphasize critical visual cues affected by challenging weather conditions, we integrated the lightweight CBAM module into our model. CBAM improves the representational capability of convolutional neural networks by applying channel and spatial attention sequentially, as shown in Figure 3. Specifically, the Channel Attention Module (M_c) identifies which features should be emphasized, while the Spatial Attention Module (M_s) determines where to focus attention spatially. Let F represent the input feature map. The operations of the channel and spatial attention modules can be mathematically expressed as:

$$M_c(F) = \sigma(MLP(AP(F)) + MLP(MP(F))) \tag{11}$$

$$M_s(F) = \sigma(f^{7\times7}([AP(F); MP(F)])) \tag{12}$$

where $AP(\cdot)$ and $MP(\cdot)$ denote average pooling and max pooling, respectively, and $\sigma(\cdot)$ represents the sigmoid function, $f^{7\times7}(\cdot)$ represents a convolution operation with a 7×7 kernel.

Then, CBAM applies the Channel Attention Module and the Spatial Attention Module sequentially as follows:

$$F' = M_c(F) \otimes F \tag{13}$$

$$F'' = M_s(F') \otimes F' \tag{14}$$

where $\otimes$ indicates element-wise multiplication. The final output represents the refined feature map, emphasizing the most discriminative and contextually relevant information for object detection [15]. Unlike standard convolution operations that blend spatial and channel features uniformly, the integration of CBAM adaptively enhances feature representations along each dimension, improving the

 doi:10.20944/preprints202511.1864.v1

overall detection performance. The motivation behind adopting CBAM lies in its ability to adaptively emphasize informative features while suppressing irrelevant or redundant ones. Adverse weather conditions such as rain streaks, fog, or snow often introduce background noise and reduce the visibility of important object cues. By sequentially applying channel and spatial attention, CBAM allows the network to focus on the most discriminative regions and feature channels, improving both localization and classification accuracy.

In our implementation, CBAM is strategically inserted after each C3k2 layer in the Neck to guide the multi-scale feature fusion process. This placement ensures that the refined features passed to subsequent layers are more semantically meaningful and less affected by degradation artifacts. Specifically, the channel attention module recalibrates the feature responses by learning inter-channel dependencies, while the spatial attention module captures spatially significant regions, highlighting target-relevant areas even under challenging lighting or visibility conditions. Together, these mechanisms enable the model to maintain high detection performance and stable feature extraction, contributing to its effectiveness in complex and adverse environments.

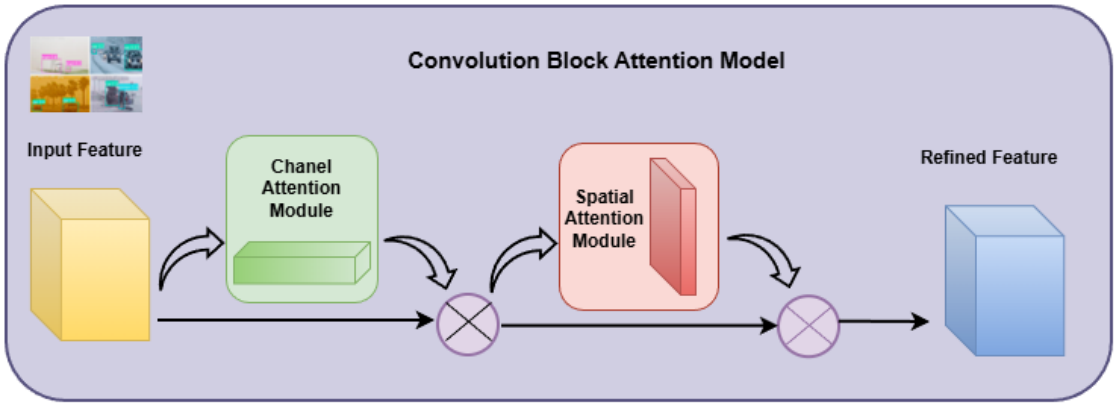

Figure 3. Structure of CBAM, consisting of two sequential attention modules: Channel and Spatial Attention.

3.5. SCDown Module

To effectively balance accuracy and efficiency, the SCDown module—originally introduced in YOLOv10 [16,17]—was utilized for its ability to decouple spatial reduction and channel compression, thereby minimizing information loss and computational complexity. The motivation behind incorporating SCDown lies in mitigating the additional overhead introduced by the TransWeather and CBAM modules, which enhance feature extraction and attention but inevitably increase parameter count and FLOPs. The SCDown module addresses this by applying two complementary convolutional branches: one focuses on spatial downsampling through strided convolution to reduce feature map resolution, while the other performs channel compression via pointwise convolution to decrease the number of feature channels. The outputs of these two branches are then fused, allowing efficient feature aggregation while maintaining semantic integrity. This design enables a more compact representation of visual features, effectively balancing lightweight computation and representational richness. By strategically replacing selected downsampling layers in the Backbone with SCDown modules, YOLOv11-TWCS achieves a significant reduction in parameters and GFLOPs, maintaining high detection accuracy and real-time inference speed even under adverse weather conditions.

4. Experiments and Results

4.1. Dataset and Experimental Setup

In this study, we chose three datasets: DAWN, Udacity, and KITTI to train and evaluate the YOLOv11-TWCS model. The DAWN dataset was used to train and evaluate our model for computer vision-based autonomous driving under adverse weather conditions [28]. Unlike most existing autonomous driving datasets, which primarily focus on normal weather, the DAWN dataset provides annotated images for four adverse weather conditions: Rainy, Snowy, Foggy, and Sandstorm. It

 doi:10.20944/preprints202511.1864.v1

contains a total of 1,020 images covering six object categories: person, bicycle, car, motorcycle, bus, and truck. Each image typically contains multiple objects, with cars being the most prevalent, followed by trucks and persons. The dataset includes extreme weather scenarios such as heavy snow, sleet, and dense fog, as well as a range of adverse weather intensities, enabling a comprehensive assessment of the model's performance across diverse conditions.

To evaluate model performance under clear weather and low-light conditions, we conducted additional experiments using the Udacity Self-Driving Car dataset. Developed for autonomous driving competitions, this dataset provides 2D bounding box annotations for consecutive video frames and includes 11 object categories: biker, car, pedestrian, trafficLight, trafficLight-Green, trafficLight-GreenLeft, trafficLight-Red, trafficLight-RedLeft, trafficLight-Yellow, trafficLight-YellowLeft, and truck. Due to the limited number of labeled instances in the trafficLight-YellowLeft category, which caused inconsistencies in performance, this class was excluded, resulting in experiments conducted on 10 object classes. The dataset contains 29,800 images at a resolution of 512×512, divided into 26,579 training images and 3,221 validation images (approximately a 9:1 split).

The KITTI dataset [29], developed by the Karlsruhe Institute of Technology (KIT) and the Toyota Technological Research Institute (TTRI), serves as a standard benchmark in autonomous driving and computer vision research. It includes 7,481 labeled training images and 7,518 test images, annotated for four object categories: Car, Cyclist, Pedestrian, and Truck. Due to the unavailability of ground-truth annotations for the test set, the training set was further divided into 3,712 training samples and 3,769 validation samples. All images were resized to a resolution of 640×640 for model input. The model was trained for 200 epochs on the KITTI dataset to assess performance under diverse urban driving conditions.

The experiments were conducted on a workstation running Windows 10, powered by an Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz and an NVIDIA GeForce RTX 2080 Ti GPU (10GB VRAM), utilizing GPU 0. The implementation used Python 3.11.0 and was built on the PyTorch 2.5.1+cu118 framework. Hardware acceleration was enabled through the CUDA platform. The detailed configuration of the model training parameters is shown in Table 1.

Table 1. Model training parameters

Parameter	Value
Epochs	300
Batch Size	8
Initial Learning Rate	0.01
Momentum	0.937
Weight Decay	0.0005
Input Size (DAWN)	416 × 416
Input Size (Udacity)	512 × 512
Input Size (KITTI)	640 × 640
Data Augmentation	Mosaic augmentation

4.2. Experimental Evaluation Criteria

To comprehensively evaluate the performance and robustness of the proposed YOLOv11-TWCS algorithm, we employ a diverse set of standard object detection metrics that jointly assess accuracy, stability, and computational efficiency. Precision, Recall, and F1-score are used to measure the trade-off between false positives and missed detections, offering insight into the model's reliability under challenging weather-induced noise. Average Precision (AP) and mean Average Precision (mAP) further quantify class-wise and overall detection quality, with both mAP@0.5 and the more rigorous mAP@0.5:0.95 reported to evaluate localization accuracy across varying IoU thresholds. In addition to accuracy metrics, we assess computational efficiency through Parameters, GFLOPs, and Model Size to determine the model's complexity and suitability for real-time deployment. Together, these evaluation indicators provide a well-rounded and meaningful assessment of YOLOv11-TWCS, ensuring that the

 doi:10.20944/preprints202511.1864.v1

model achieves a balanced combination of high accuracy and operational efficiency for autonomous driving in adverse weather conditions.

4.3. Performance Analysis on DAWN Dataset

In this subsection, we evaluate the generalization ability of our proposed model across a wide range of scenarios. We trained and validated the YOLOv11-TWCS model on the DAWN dataset, which includes four challenging adverse weather conditions: Rainy, Snowy, Foggy, and Sandstorm. Additionally, we compared its performance with baseline methods, other commonly used YOLO series algorithms, and state-of-the-art models. The experimental results are summarized in Table 2 and Figure 4.

Table 2 demonstrates the efficiency and effectiveness of different models in terms of parameters, GFLOPs, precision, recall, mAP, and FPS. Traditional two-stage detectors such as Faster R-CNN require high computational resources (138M parameters, 231.7 GFLOPs) but achieve limited accuracy (26.4% mAP@0.5) and slow inference (29 FPS). Single-stage detectors like RT-DETR V1 and V2 [22,23], reduce computation but only show moderate improvements in mAP while maintaining real-time speed (30 FPS).

The YOLO series models show progressive improvements in both detection accuracy and speed. YOLOv6m, YOLOv8m, and YOLOv9m achieve higher mAP@0.5 (44.2%–48.5%) with FPS ranging from 72 to 98. YOLOv11 and YOLOv12m further improve precision and mAP while sustaining high inference speeds. The proposed YOLOv11-TWCS model surpasses all compared models, achieving the highest precision (57.2%) and recall (64.1%), leading mAP@0.5 (59.1%) and mAP@0.5:0.95 (36.1%), while remaining compact (19.32M parameters) and highly efficient (105 FPS). These results demonstrate that YOLOv11-TWCS effectively balances accuracy, speed, and computational cost, confirming its robustness and generalization ability on the DAWN dataset across diverse adverse weather conditions.

Table 2. Comparative experimental results on the DAWN dataset with state-of-the-art models

Models	Params (M)	GFLOPs	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
SwinTransformer [21]	51.3	455	56.4	50.8	46.2	27.73	27.73
YOLOv9m	20.0	76.9	49.7	50.8	48.5	22.8	98
RT-DETR V1 [22]	20.09	61.17	-	-	37.70	21.80	30.39
YOLOv6m	51.9	161	47.9	46.9	44.2	28.5	72
SPR-YOLO [25]	3.9	14.6	63.5	33.5	50.8	30.48	214.68
Faster R-CNN [4]	138	231.7	53.04	56.5	26.4	14.01	29
PVDM-YOLOv8l [26]	104.7	458	75.8	50.4	55.2	30.2	70.0
RT-DETR V2 [23]	20.09	61.17	-	-	36.46	21.98	30.67
YOLOv8_MobileNetV3 [25]	2.3	5.7	51.2	45.6	40.4	24.6	162.48
YOLOv8m	25.8	79	49.3	50.7	47.7	22.0	97
YOLOv13	27.57	89.0	45.1	43.0	49.0	21.0	91
YOLOv11m	23.88	85.5	56.0	44.2	57.9	35.7	93
YOLOv12m	20.1	67.1	54.4	45.6	54.8	33.2	105
YOLOv11-TWCS	**19.32**	**67.5**	**57.2**	**64.1**	**59.1**	**36.1**	**105**

 doi:10.20944/preprints202511.1864.v1

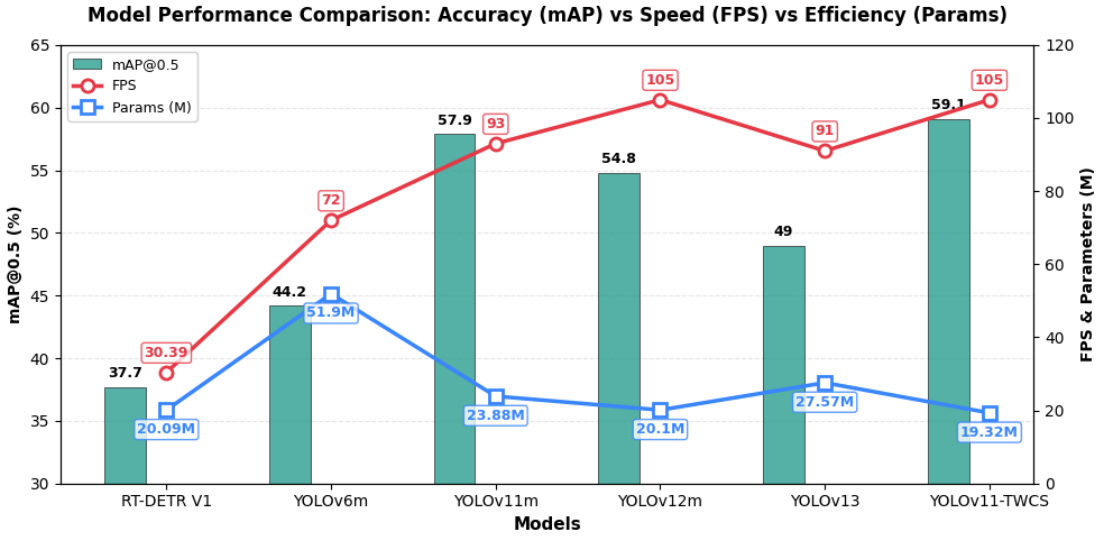

Figure 4. Evaluation of model detection accuracy, processing speed, and computational efficiency on the DAWN dataset

4.4. Ablation Study

To evaluate the contribution of each proposed module in YOLOv11-TWCS, we conducted an ablation study on DAWN dataset, as summarized in Table 3. The baseline YOLOv11m model achieves a mAP@0.5 of 57.9% and mAP@0.5:0.95 of 35.7%, with 23.83M parameters, 85.5 GFLOPs, and 93 FPS. This establishes the reference point for assessing the effectiveness of the TransWeather, CBAM, and SCDown modules when integrated individually or in combination. Adding the TransWeather module to the baseline (Model 2) improves mAP@0.5 to 58.5% and mAP@0.5:0.95 to 35.9%, highlighting the module's effectiveness in capturing global and local features under adverse weather conditions. Interestingly, this modification also reduces the parameter count to 21.8M and slightly increases FPS to 100, indicating improved computational efficiency likely due to optimized feature extraction. Similarly, incorporating CBAM alone (Model 3) achieves a mAP@0.5 of 58% and mAP@0.5:0.95 of 36%, demonstrating the benefit of attention mechanisms in emphasizing salient features, though FPS slightly decreases to 91 due to additional computations.

Integrating SCDown alone (Model 4) improves FPS to 97 while maintaining comparable accuracy (mAP@0.5: 58%, mAP@0.5:0.95: 35.8%), showing its efficiency in reducing computation. Combining modules further boosts performance: Model 5 (Weather Transformer and CBAM) achieves 58.9% mAP@0.5, 36% mAP@0.5:0.95, 20.51M parameters, and 103 FPS. The full YOLOv11-TWCS (Model 6), with all three modules, delivers the best results—59.1% mAP@0.5, 36.1% mAP@0.5:0.95, lowest parameters (19.32M), reduced GFLOPs (67.5), and fastest inference (105 FPS). These findings highlight that TransWeather enhances feature extraction under adverse conditions, CBAM emphasizes key regions, and SCDown reduces computation, with their integration achieving a balanced, high-speed, and accurate detection model.

Table 3. Ablation study analyzing the contribution of individual components in YOLOv11-TWCS

Models	**Components**				**mAP@0.5**	**mAP@0.5:0.95**	**Params (M)**	**FPS**
	Baseline	**ransWeather**	**CBAM**	**SCDown**				
Model 1	✓				57.9	35.7	23.83	93
Model 2	✓	✓			58.5	35.9	21.80	100
Model 3	✓		✓		58.0	36.0	24.10	91
Model 4	✓			✓	58.0	35.8	22.00	97
Model 5	✓	✓	✓		58.9	36.0	20.51	103
Model 6	✓	✓	✓	✓	**59.1**	**36.1**	**19.32**	**105**

 doi:10.20944/preprints202511.1864.v1

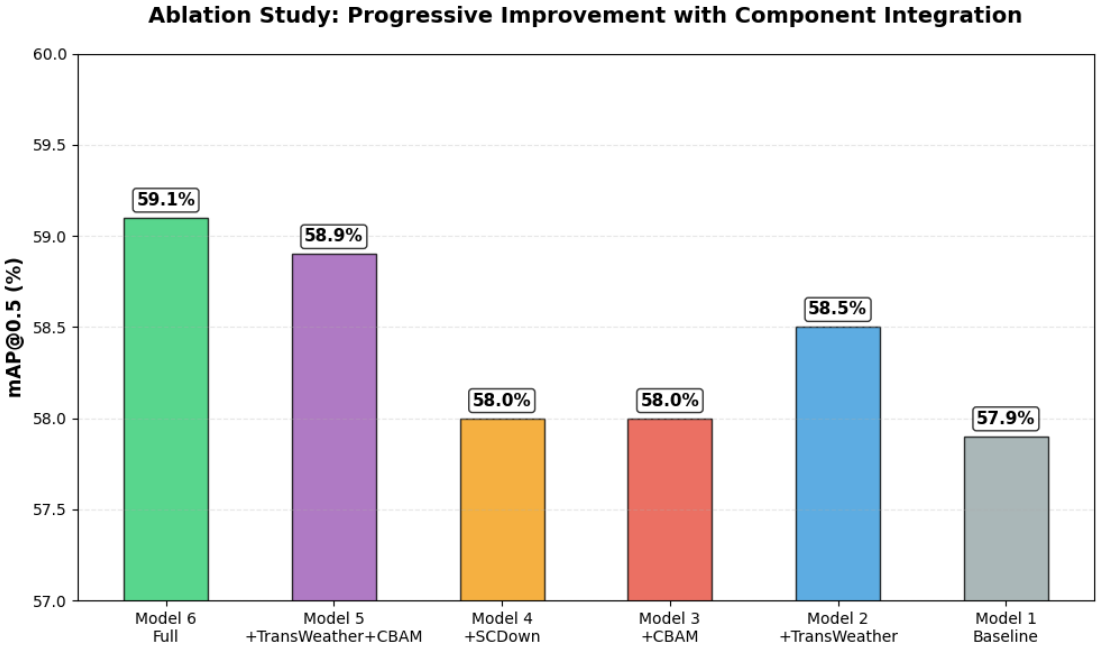

Figure 5. Contribution of Each Module to YOLOv11-TWCS Performance

4.5. Generalization Performance on Udacity and KITTI Datasets

In this subsection, we evaluate the generalization ability of our proposed model across a wider range of scenarios. To this end, we consider not only the DAWN dataset, as adopted in the previous section, but also the Udacity Self-Driving Car dataset and the KITTI dataset.

The Udacity dataset, developed for autonomous driving competitions, includes 2D bounding box annotations over video sequences and covers 11 object categories under clear weather and low-light conditions. The KITTI dataset, created by the Karlsruhe Institute of Technology (KIT) and Toyota Technological Research Institute (TTRI), provides over 15,000 annotated images for four categories: Car, Cyclist, Pedestrian, and Truck. Together, these datasets serve as complementary benchmarks to evaluate performance across adverse weather, illumination variation, and complex urban driving conditions.

As shown in Table 4, classical detectors such as SSD and Faster R-CNN achieve limited performance. SSD reports low accuracy on both KITTI (50% mAP@0.5, 18% mAP@0.5:0.95) and Udacity (52.1% mAP@0.5, 20.3% mAP@0.5:0.95), with modest speeds of around 22–28 FPS. Faster R-CNN improves accuracy (72.5% mAP@0.5 on KITTI, 78.6% on Udacity) but is computationally expensive, with 138M parameters and 183 GFLOPs, and operates at just 11 FPS, making it impractical for real-time applications. In contrast, YOLO-based detectors provide a better trade-off. YOLOv4 and TPH-YOLOv5 improve accuracy while maintaining faster speeds, with TPH-YOLOv5 achieving 79.5% mAP@0.5 on KITTI and 86.5% on Udacity, along with high mAP@0.5:0.95 values and over 35 FPS.

Among the most recent YOLO versions, YOLOv11 achieves strong results with 80% mAP@0.5 on KITTI and 84.7% on Udacity, while YOLOv12 and YOLOv13 remain competitive but slightly lower. The proposed YOLOv11-TWCS model surpasses all baselines, achieving the best overall performance. On KITTI, it records 81.9% mAP@0.5 and 56.9% mAP@0.5:0.95, with a speed of 245 FPS. On Udacity, it further improves to 88.5% mAP@0.5 and 58.1% mAP@0.5:0.95, while sustaining 146 FPS. Moreover, it achieves these results with the lowest parameter count (19.32M) and reduced GFLOPs (67.5), demonstrating superior efficiency. These findings confirm that YOLOv11-TWCS generalizes well across diverse datasets, consistently balancing accuracy, computational efficiency, and real-time performance.

 doi:10.20944/preprints202511.1864.v1

Table 4. Comparative experimental results on the Udacity and KITTI datasets with state-of-the-art models

Model	KITTI dataset			Udacity dataset			Params	GFLOPs
	mAP50	mAP50-95	FPS	mAP50	mAP50-95	FPS	(M)	
YOLOv4 [18]	77.6	46.1	48	83.4	50.2	50.2	27.6	26
TPH-YOLOv5 [20]	79.5	51.2	35.5	86.5	56.5	36.3	95	237
RFBNet [19]	73.44	39.20	32.5	77.9	41.2	32	34.5	95.0
SSD [6,7]	50	18	28	52.1	20.3	21.8	24	31.4
Faster R-CNN [4]	72.5	45.5	11	78.6	49.4	10.7	138	183
RTMDet [24]	80.0	48.0	180	84.8	50.4	178	16	27.7
YOLOv11	80	55	227.3	84.7	56	135	23.88	85.5
YOLOv12	79.1	54.1	220	83.1	55	130	20.1	67.1
YOLOv13	78.5	52	215	82	54	125	27.57	89.0
YOLOv11-TWCS	**81.9**	**56.9**	**245.1**	**88.5**	**58.1**	**146**	**19.32**	**67.5**

4.6. Visualization Results on DAWN Dataset

This subsection evaluates leading object detection models under four challenging weather conditions—sandstorm, foggy, snowy, and rainy—using the DAWN dataset. These conditions test model robustness against reduced visibility, occlusion, and reflections. Figure 6 shows detection results, highlighting each model's stability, accuracy, and reliability in adverse real-world driving scenarios.

4.6.1. Sandstorm Condition Analysis

Under sandstorm conditions, illustrated in the first column of Figure 6, the visual contrast is drastically reduced due to suspended dust particles, severely limiting object visibility. YOLOv8m struggled in this setting, detecting only the closest vehicle and completely failing to identify the second and third vehicles in the scene. These false negatives highlight the model's limited capacity to extract discriminative features under particulate-heavy environments. Both YOLOv12 and YOLOv13 achieved partial improvements, detecting two vehicles but still missing distant targets, confirming that visibility degradation continues to hinder accurate long-range detection.

YOLOv11 exhibited stronger feature extraction performance, successfully detecting all three vehicles, though its detection confidence for the middle car remained relatively low (0.3), suggesting uncertainty under dense sand occlusion. In contrast, the proposed YOLOv11-TWCS model achieved outstanding robustness, accurately detecting all vehicles with an average confidence of 0.89. This performance improvement is primarily attributed to the TransWeather block, which mitigates the scattering and contrast loss effects induced by airborne particles, while the CBAM attention mechanism enhances the model's focus on salient features, maintaining accurate spatial awareness even under low visibility.

4.6.2. Foggy Weather Evaluation

Fog conditions, shown in the second column of Figure 6, introduce heavy occlusion and blur, leading to frequent misclassification in standard models. YOLOv8m performed poorly in this scenario, generating duplicate detections of the same truck and incorrectly labeling parts of the truck as separate vehicles—severe false positives that compromise scene interpretation. YOLOv12 partially improved

 doi:10.20944/preprints202511.1864.v1

object localization but still exhibited duplicate bounding boxes and inconsistent category predictions, alternately classifying the truck as both a "truck" and a "person."

Both YOLOv11 and YOLOv13 correctly detected the two trucks but suffered from reduced confidence scores, indicating weaker feature certainty in low-contrast fog conditions. In contrast, YOLOv11-TWCS achieved consistently accurate detection, correctly identifying both trucks with average confidence scores and without duplication or misclassification.

4.6.3. Snowy Weather Detection Challenges

The snowy condition, depicted in the third column, represents one of the most challenging environments due to uniform backgrounds and reflective snow surfaces that obscure object edges. YOLOv8m correctly detected most vehicles but mistakenly classified one car as a truck, showing difficulty in distinguishing object contours. YOLOv12 performed worse, failing to detect all intermediate vehicles, introducing dangerous false negatives that could critically impact automated decision-making. YOLOv11 and YOLOv13 detected the vehicles correctly but introduced redundant bounding boxes for the leading car, indicating confusion in object boundary separation. Such redundancy complicates trajectory estimation and increases post-processing costs. The proposed YOLOv11-TWCS, however, achieved perfect detections with no duplicates or misclassifications. This precision is largely due to the Spatial-Channel Decoupled Downsampling (SCDown) mechanism, which preserves spatial consistency and texture gradients even when global brightness dominates the scene.

4.6.4. Rainy Condition Performance

Rainy conditions create dynamic distortions such as water streaks, blurred reflections, and varying brightness. YOLOv11-TWCS excelled in this scenario, accurately detecting both the vehicle and pedestrian without any missed detections, false positives, or duplicate predictions.

In contrast, YOLOv8m missed the nearby vehicle entirely, detecting only the pedestrian, representing a serious false negative. YOLOv11 and YOLOv12 identified both objects but produced redundant bounding boxes, reducing interpretability and potentially triggering false alarms. YOLOv13 showed instability, initially misclassifying the car as a truck before correction. The proposed YOLOv11-TWCS maintained consistent performance due to its adaptive attention mechanisms and weather-invariant feature extraction, ensuring reliable detection despite reflections and fluctuating luminance.

YOLOv11-TWCS delivers stable and accurate performance across all weather conditions, minimizing false negatives and false positives while maintaining high confidence and real-time speed. The TransWeather block enhances feature robustness under visual degradation, the CBAM module focuses attention on critical regions for precise detection, and the SCDown module improves efficiency by preserving key spatial details with reduced computation, creating a compact and high-performing model suitable for real-world autonomous driving.

 doi:10.20944/preprints202511.1864.v1

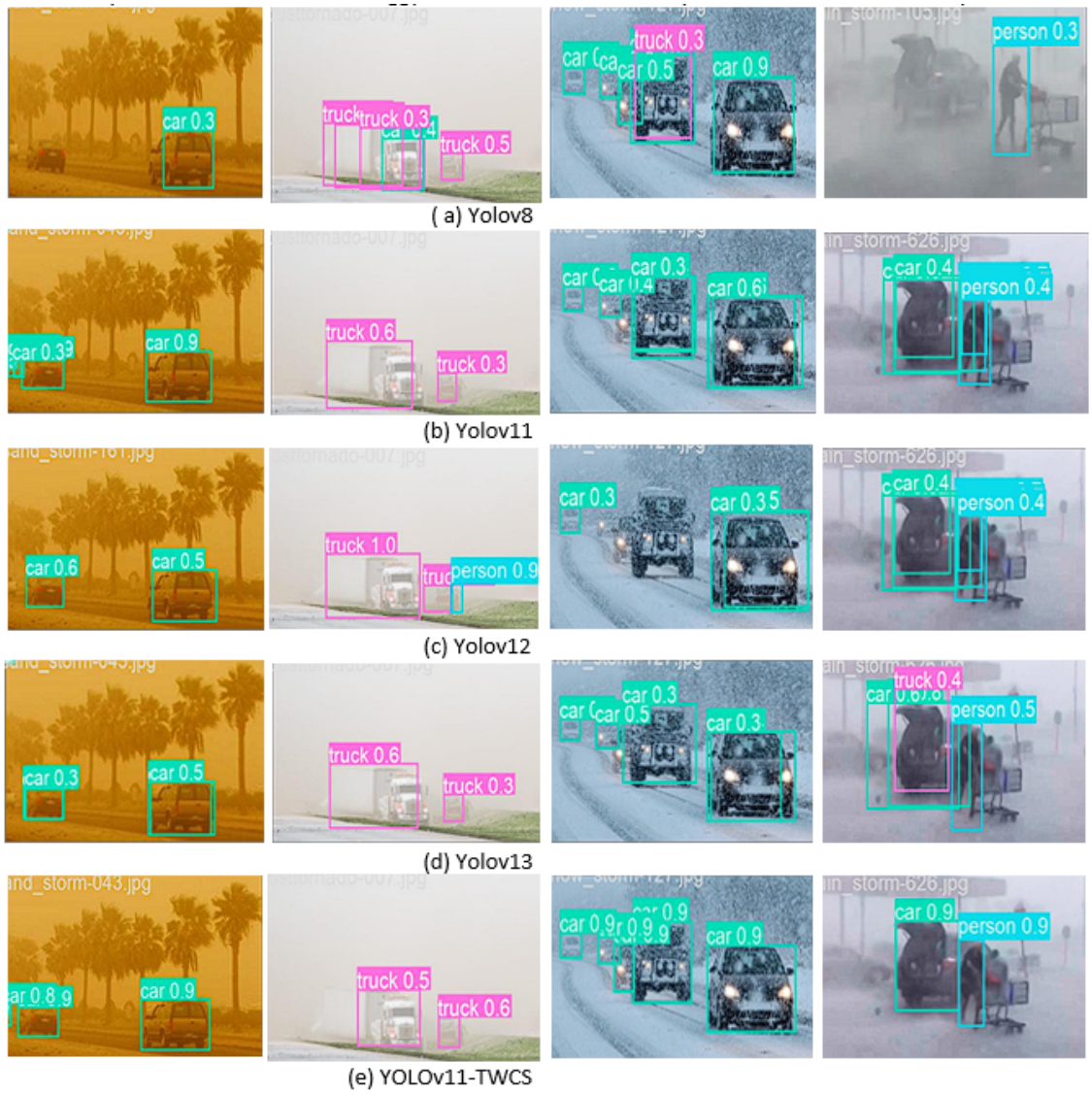

Figure 6. Comparison of object detection results under adverse weather conditions for different YOLO models: (a) YOLOv8, (b) YOLOv11, (c) YOLOv12, (d) YOLOv13, and (e) YOLOv11-TWCS (ours). Each column within the sets represents a specific weather condition from left to right: sandstorm, fog, snow, and rain.

5. Discussion

The experimental results demonstrate that YOLOv11-TWCS significantly outperforms existing state-of-the-art models across multiple datasets and adverse weather conditions. The integration of TransWeather, CBAM, and SCDown modules creates a synergistic effect that enhances both detection accuracy and computational efficiency.

The TransWeather block proves particularly effective in handling weather-induced distortions by dynamically adjusting attention to brightness variations and noise patterns. This capability is crucial for maintaining detection performance in conditions where traditional models struggle, such as sandstorms and heavy fog where contrast is severely reduced.

The CBAM module contributes to improved feature discrimination by selectively emphasizing relevant spatial regions and channel features. This attention mechanism helps the model focus on critical object characteristics while suppressing weather-induced noise, leading to higher precision and recall rates.

The SCDown module successfully mitigates the computational overhead introduced by the attention mechanisms, demonstrating that efficiency improvements can be achieved without sacrificing

 doi:10.20944/preprints202511.1864.v1

detection accuracy. This balance is essential for real-time autonomous driving applications where both speed and reliability are paramount.

The consistent performance improvement across DAWN, KITTI, and Udacity datasets indicates that YOLOv11-TWCS possesses strong generalization capabilities. This robustness across different environmental conditions and dataset characteristics suggests that the proposed architecture effectively addresses fundamental challenges in adverse weather object detection.

6. Conclusions

In this paper, we introduced YOLOv11-TWCS, an enhanced object detection model developed to improve performance in adverse weather and data-limited environments while maintaining high computational efficiency. The model integrates the TransWeather block into the C2PSA module of the YOLOv11 backbone to enable weather-adaptive feature extraction, while the CBAM module is incorporated after each C3k2 layer in the Neck to highlight essential features and suppress irrelevant ones. To further optimize efficiency, certain convolutional layers were replaced with SCDown modules, reducing parameters and GFLOPs while boosting FPS. Comprehensive experiments on the DAWN, Udacity (clear and low-light), and KITTI datasets demonstrate that YOLOv11-TWCS delivers significant improvements in detection accuracy, efficiency, and generalization across diverse weather and lighting conditions. Overall, the proposed model offers a robust and lightweight solution for real-time autonomous driving applications, with future work focusing on addressing limitations in dense or cluttered scenes and extending adaptability to more extreme weather scenarios for enhanced reliability and safety.

Author Contributions: Conceptualization, C.M. and H.W.; methodology, C.M and H.W.; investigation, C.M.; writing—original draft preparation, C.M.; writing—review and editing, H.W.; supervision, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this study (DAWN, KITTI, and Udacity) are publicly available.

Acknowledgments: The authors would like to thank Donghua University for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

1. R. Zhao, S. H. Tang, J. Shen, E. E. B. Supeni, and S. A. Rahim, "Enhancing autonomous driving safety: A robust traffic sign detection and recognition model TSD-YOLO," *Signal Process.*, vol. 225, Dec. 2024, Art. no. 109619.
2. D. Dai, J. Wang, Z. Chen, and H. Zhao, "Image guidance based 3D vehicle detection in traffic scene," *Neurocomputing*, vol. 428, pp. 1–11, Mar. 2021.
3. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
4. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE TPAMI*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
5. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE TPAMI*, vol. 42, no. 2, pp. 386–397, Feb. 2020.
6. W. Liu et al., "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.
7. L. Zheng, C. Fu, and Y. Zhao, "Extend the shallow part of single shot multibox detector via convolutional neural network," *Proc. SPIE*, vol. 10806, Aug. 2018.
8. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun. 2016, pp. 779–788.
9. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. CVPR*, Jul. 2017, pp. 6517–6525.
10. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
11. W. Chu et al., "Multi-task vehicle detection with region-of-interest voting," *IEEE Trans. Image Process.*, 2017.

12. R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," *arXiv preprint*, arXiv:2410.17725, 2024.
13. S. Sun, W. Ren, X. Gao, R. Wang, and X. Cao, "Restoring Images in Adverse Weather Conditions via Histogram Transformer," *arXiv preprint*, arXiv:2407.10172, Jul. 2024.
14. Valanarasu, J.M.J.; Yasarla, R.; Patel, V.M. TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions. 2024, arXiv:2409.06334.
15. S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, Munich, 2018, pp. 3-19.
16. D. Chen and L. Zhang, "SL-YOLO: A stronger and lighter drone target detection model," 2024, arXiv:2411.11477.
17. A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," 2024, arXiv:2405.14458.
18. A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao,"YOLOv4: Optimal speed and accuracy of object detection," 2020, arXiv:2004.10934.
19. S. Liu, D. Huang, and Y. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," arXiv preprint arXiv:1711.07767, Jul. 2018.
20. X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.
21. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021
22. Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y.Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.
23. W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, and Y. Liu, "RT DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer," 2024, arXiv:2407.17140.
24. C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: An Empirical Study of Designing Real-Time Object Detectors," arXiv preprint arXiv:2212.07784, Dec. 16, 2022.
25. H. Liu, Y. Ma, H. Jiang, and T. Hong, "SPR-YOLO: A Traffic Flow Detection Algorithm for Fuzzy Scenarios," *Arabian Journal for Science and Engineering*, King Fahd University of Petroleum and Minerals, accepted Jan. 2025.
26. N. U. A. Tahir, Z. Zhang, M. Asim, S. Iftikhar, and A. A. Abd El-Latif, "PVDM-YOLOv8l: A solution for reliable pedestrian and vehicle detection in autonomous vehicles under adverse weather conditions," *Multimedia Tools and Applications*, Springer Nature, accepted Sept. 2024.
27. Z. Su, J. Yu, H. Tan, X. Wan, and K. Qi, "MSA-YOLO: A remote sensing object detection model based on multi-scale strip attention," *Sensors*, vol. 23, no. 15, p. 6811, Jul. 2023.
28. M. A. Kenk and M. Hassaballah, "DAWN: Vehicle detection in adverse weather nature dataset," 2020, arXiv:2008.05402.
29. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.