DW-YOLO: A Robust Object Detection Framework for Adverse Weather and Domain Shifts

Shuo Cai[1], Shuai Chen[2], Yuanzhi Tang[2], Wei Shuai[1], Zeyang Deng[2], Hao Xiong[2]
[1]School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha, China
[2]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China

***Abstract*—Under adverse environmental conditions such as fog, rain, snow, and low illumination, object detection models often suffer significant performance degradation in real-world deployments due to domain shifts and visual feature degradation. To address this problem, this paper adopts YOLOv8 as the baseline detector and proposes a robust framework, termed DW-YOLO, which is designed to learn domain-invariant object representations in complex environments.At the core of the framework is the proposed Dual-Context Attention Module (DCAM). This module integrates short-distance attention and long-distance attention with multi-scale contextual fusion, enabling simultaneous modeling of global contextual relationships and local fine-grained details. By explicitly leveraging complementary contextual information, DCAM substantially improves the robustness of the characteristics against weather-induced domain drift. Furthermore, an adaptive optimization strategy, Weighted Intersection over Union (WIoU), is incorporated to emphasize challenging and underrepresented samples during training. This strategy improves generalization across diverse environmental conditions while preserving an end-to-end training pipeline and without incurring significant computational overhead. Extensive experiments on multiple public benchmark datasets demonstrate that DW-YOLO consistently outperforms baseline detectors under adverse weather conditions, yielding notable improvements in both detection accuracy and robustness. These findings indicate that DW-YOLO provides an effective solution to object detection under domain shift and shows strong potential for deployment in real-world intelligent perception systems.**

***Index Terms*—Object Detection, YOLO, Computer Vision, Image Processing, Domain Shifts.**

I. Introduction

Object detection has become a fundamental component in modern perception systems, supporting applications such as autonomous driving, intelligent transportation, and video surveillance. With the rapid development of deep convolutional networks, detection performance has advanced from two-stage frameworks such as Faster R-CNN [1] to efficient one-stage detectors including SSD [2]. Among them, the YOLO family [3]–[6] has gained significant attention due to its favorable balance between accuracy and efficiency, making it suitable for real-time deployment.

However, when detectors are deployed in real-world environments, performance often degrades severely under adverse weather conditions such as fog, rain, snow, and low illumination. Visibility degradation leads to feature attenuation, contrast reduction, and noise amplification, resulting in unstable localization and missed detections. Moreover, such degradations introduce distribution discrepancies between training data and deployment scenarios, giving rise to the well-known problem of *domain shift*. Models trained mainly on clear-weather datasets may therefore generalize poorly to adverse-weather domains.

Existing research has attempted to mitigate these issues along two main directions. The first focuses on architectural improvements. Recent YOLO variants [4]–[6] enhance feature extraction and optimization strategies while maintaining real-time capability, yet most are still optimized on clear images and remain sensitive to visibility degradation. The second emphasizes robustness mechanisms, including image enhancement, contextual modeling, and domain-oriented optimization. Dehazing-based preprocessing [7], [8] can improve visibility but may introduce artifacts that propagate into detectors. Context-aware networks [9]–[11] strengthen multi-scale and long-range representations, while domain adaptation approaches [12]–[14] alleviate distribution mismatch across domains. Despite these advances, integrating such mechanisms into a unified and lightweight detector remains challenging, particularly under complex weather-induced degradations.

Motivated by these challenges, this paper proposes DW-YOLO, a robust detection framework tailored for adverse-weather and cross-domain scenarios. Built upon YOLOv8 [5], DW-YOLO enhances both representation learning and localization robustness. The proposed DCAM jointly exploits local structural cues and global contextual dependencies, effectively alleviating feature attenuation under low-contrast conditions. In addition, a weighted IoU regression objective inspired by WIoU [15] stabilizes bounding-box optimization, while diversified data-augmentation strategies further enhance cross-domain generalization.

Extensive experiments on representative adverse-weather benchmarks, including FOG-Driving [16], RTTS [17], and Foggy Cityscapes [16], demonstrate that DW-YOLO consistently outperforms baseline YOLO models and recent robust detectors, achieving improved accuracy and generalization across synthetic and real fog domains.

The main contributions of this work are summarized as follows:

- We propose DW-YOLO, a unified and lightweight detection framework designed for adverse-weather and cross-domain conditions.
- We design a Dual-Context Attention Module that enhances both long-range dependency modeling and fine-

Corresponding author: Shuo Cai (caishuo@csust.edu.cn)

grained perception under feature attenuation.
- We incorporate a WIoU-based adaptive localization objective together with diversification-oriented augmentation to improve robustness against domain shifts.
- The proposed DW-YOLO was comprehensively evaluated across multiple datasets, demonstrating consistent performance improvements over baseline methods.

II. Related Work

A. Object Detection under Adverse Weather

Adverse weather such as fog, rain, snow, and low illumination introduces scattering, blur, and contrast reduction, which weakens visual features and leads to noticeable drops in detection accuracy. Early work primarily relied on image-level restoration before detection. Classical priors such as the dark channel prior modeled atmospheric scattering and achieved promising results in dehazing [7]. Subsequent CNN-based approaches [8], [18]–[20] improved visual restoration through learned feature fusion and multi-scale inference. However, these methods mainly optimize human-perceived clarity rather than detection objectives, and may introduce artifacts that negatively affect bounding box regression.

To alleviate such task misalignment, researchers began integrating enhancement and detection into unified frameworks. Image-adaptive architectures such as IA-YOLO [21] and ERUP-YOLO [22] embed learnable preprocessing into the detector, enabling dynamic adaptation to visibility degradation. Multi-scale and cross-modal approaches such as MASFNet [23] and CME-YOLO [24] further exploit complementary cues to stabilize perception in foggy scenes.

A second direction improves robustness directly within detection architectures. Works including R-YOLO [25], D-YOLO [26], and TTSDA-YOLO [27] introduce weather-aware modules and specialized optimization strategies to reduce performance collapse under degradation. Despite these advances, many methods remain closely tied to specific enhancement pipelines or incur additional computational cost, limiting scalability across unseen environments.

These challenges motivate the need for a unified and lightweight framework that enhances representation robustness and localization stability, while avoiding dependence on task-misaligned restoration.

B. Context-Aware and Domain-Robust Mechanisms

Contextual information is essential for recognizing partially occluded or low-contrast objects. Feature pyramid networks [9] enhance multi-scale representation, while non-local networks [10] and squeeze-and-excitation modules [11] strengthen long-range dependency modeling and channel reweighting. Transformer-based detectors [28] further introduce global reasoning, albeit often at notable computational cost.

Beyond context modeling, *domain shift* remains a major obstacle when transferring models from clear-weather training data to adverse-weather deployment. Domain-adaptive detectors such as DA-Faster R-CNN [12] and domain-oriented YOLO extensions [13], [14] align feature distributions between domains. More recent works explore auxiliary domain guidance [29], pseudo-label refinement [30], and domain generalization strategies [31]. However, integrating domain robustness and lightweight context modeling into a unified framework remains insufficiently explored.

Motivated by these findings, our work combines dual-context attention with adaptive localization and data-driven regularization, aiming to simultaneously enhance feature robustness and cross-domain stability.

III. Proposed Method

A. Our Model Architecture

We propose DW-YOLO, a robust object detection framework designed to cope with adverse weather conditions and domain shifts, as illustrated in Fig. 1.

The model is built upon YOLOv8 and follows the standard single-stage pipeline composed of a backbone, neck, and detection head, while introducing targeted enhancements to improve robustness and generalization.

To alleviate feature degradation caused by fog, rain, and low illumination, DCAMs are integrated into multiple stages of the backbone. By simultaneously capturing local contextual details and long-range dependencies, DCAM strengthens the discriminability of low-contrast features and enables more reliable perception in visually degraded environments.The neck adopts a feature-pyramid fusion structure that aggregates DCAM-enhanced features across multiple spatial resolutions. This design allows the detector to better handle objects with different sizes and visibility levels, which is particularly important in adverse weather scenarios.In the detection head, the WIoU is employed to improve bounding-box regression stability. WIoU adaptively focuses on informative and moderately hard samples while suppressing unreliable gradients from noisy predictions. In addition, diversified data augmentation is applied during training to expand the data distribution and reduce the effect of domain shifts.

Overall, DW-YOLO jointly improves feature representation, optimization stability, and data-level robustness, resulting in stronger detection performance under both adverse weather and cross-domain conditions.

B. Dual Context Attention Module

To address the limitations of contextual modeling in traditional YOLO variants under complex weather conditions, we propose the Dual Context Attention Module (DCAM), whose structure is shown in Fig. 2. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, DCAM first extracts local contextual cues through a lightweight local operator, forming the Short-Distance Attention (SDA) branch. This process is formulated as

$$L = f_{\text{local}}(X), \tag{1}$$

where $f_{\text{local}}(\cdot)$ denotes a stack of depthwise convolution and normalization operations.

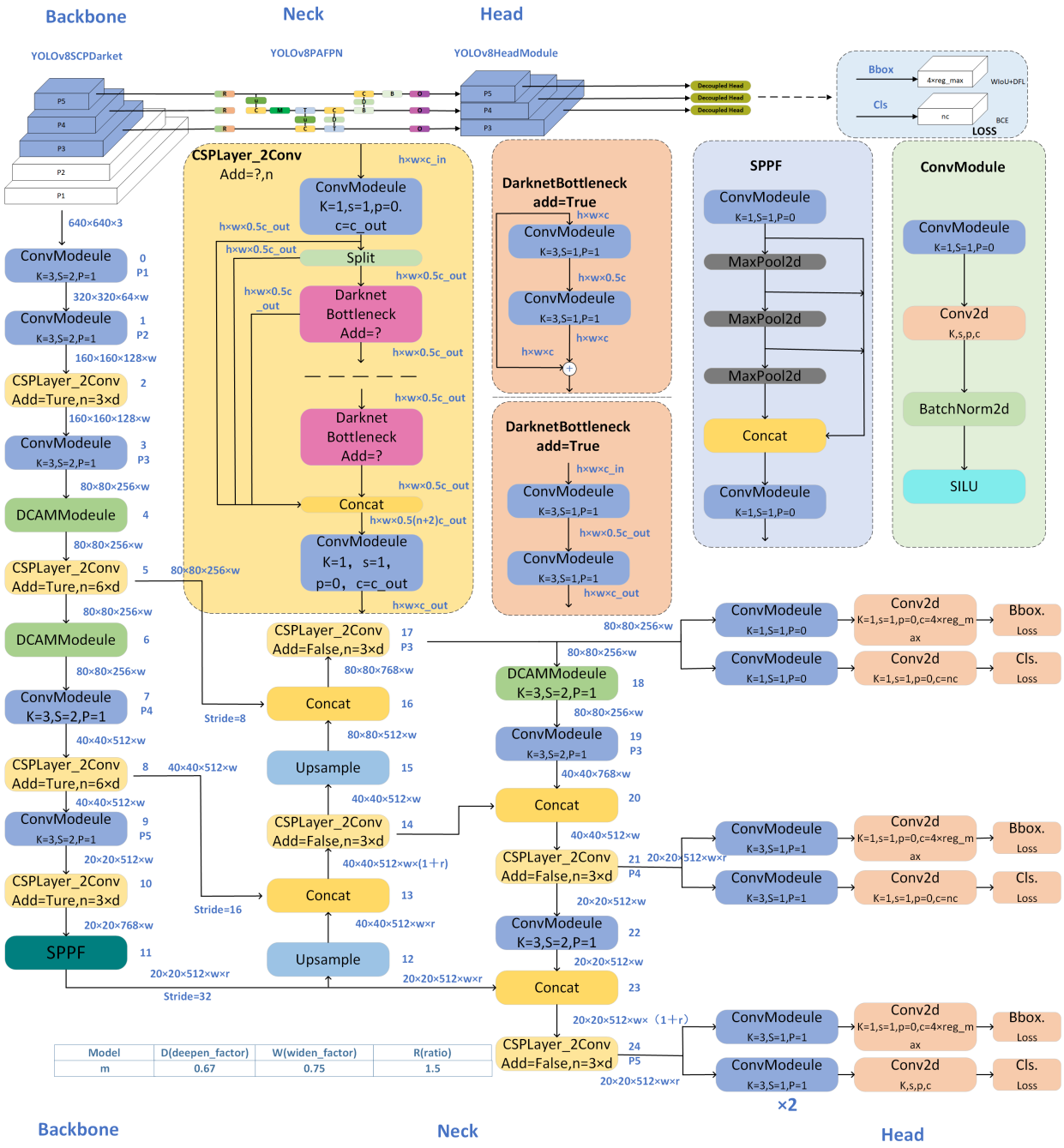


Fig. 1. Overall architecture of the proposed model.

Meanwhile, DCAM constructs a Long-Distance Attention (LDA) branch to capture global dependencies. The input feature is projected into Query, Key, and Value spaces via 1×1 convolutions:

$$Q = W_Q * X, \quad K = W_K * X, \quad V = W_V * X, \tag{2}$$

where W_Q, W_K, and W_V denote learnable projection kernels.

The attention score matrix is computed through similarity matching between Q and K:

$$A = \mathrm{Softmax}(Q^\top K). \tag{3}$$

Using the attention map A, the global enhanced feature is obtained by

$$V' = AV. \tag{4}$$

To fuse local geometric cues with global semantic interactions, DCAM concatenates V' and L channel-wise and applies

a 1×1 convolution for feature fusion:

$$F = \phi\left(\mathrm{Concat}(V', L)\right), \tag{5}$$

where $\phi(\cdot)$ denotes the fusion transformation.

Finally, a residual connection is applied to produce the output:

$$Y = X + F. \tag{6}$$

Through the integration of local context modeling (SDA) and global dependency modeling (LDA), DCAM jointly exploits static spatial structures and dynamic long-range interactions, thereby improving robustness and accuracy in low-contrast scenarios such as fog, rain, snow, and low illumination.

C. WIoU-Based Optimization with Data Augmentation

Object detection under adverse weather conditions often suffers from inaccurate localization and unstable training due to visibility degradation and domain shift. Traditional IoU-based losses provide limited robustness. IoU focuses only on overlap and ignores geometric relations. DIoU and CIoU introduce center-distance and aspect-ratio constraints to improve regression stability [32], [33], but they still treat all samples equally during optimization.

In this work, we adopt WIoU as the bounding box regression loss [15]. WIoU introduces a dynamic focusing mechanism that suppresses easy, high-quality samples while assigning more optimization attention to medium- and hard-quality predictions. As illustrated in Fig. 3, WIoU produces smoother and more discriminative regression curves compared with IoU, DIoU, and CIoU, making it more suitable for foggy and low-contrast scenes. This leads to more stable training and more accurate localization than conventional IoU-based losses.

WIoU introduces a dynamic focusing mechanism that adaptively adjusts the contribution of each training sample based on its localization quality. Unlike conventional IoU-based losses, which treat all samples equally, WIoU assigns higher weights to hard samples while suppressing the influence of low-quality or noisy predictions.In adverse weather conditions, object boundaries are often blurred and localization uncertainty increases. Under such circumstances, conventional IoU-based losses tend to be dominated by noisy or unreliable samples. WIoU alleviates this issue by adaptively down-weighting low-quality predictions and emphasizing moderately hard samples, which leads to more stable optimization under degraded visual conditions. The WIoU loss can be formulated as:

$$L_{\mathrm{WIoU}} = 1 - \mathrm{IoU}\left(B_p, B_{gt}\right) + \lambda \cdot w\left(B_p, B_{gt}\right) \tag{7}$$

Here, B_p and B_{gt} denote the predicted bounding box and the ground truth bounding box, respectively; $\mathrm{IoU}\left(B_p, B_{gt}\right)$ represents the Intersection over Union; λ is the balance coefficient, used to control the influence of the weighting term; $w\left(B_p, B_{gt}\right)$ is the difficulty weighting function, which adapts the optimization intensity based on sample characteristics.

In adverse weather scenarios, visual degradations such as fog, rain, and low illumination significantly alter object appearance and introduce large intra-class variations. To improve generalization under such domain shifts, we incorporate a set of data augmentation techniques during training, including photometric distortion, random scaling, spatial flipping, and weather-related perturbations [3], [34]. These augmentations expose the model to diverse visual conditions and object scales, encouraging the learned representations to be invariant to environmental changes.

The combination of WIoU-based optimization and data augmentation yields a complementary effect. While data augmentation expands the diversity of training samples and simulates adverse conditions, WIoU dynamically rebalances the learning process by focusing on informative samples with reliable gradients. Together, they promote stable convergence and improve localization accuracy, particularly for small, occluded, and low-contrast objects commonly encountered in adverse weather environments.

IV. Experimental Results

A. Dataset

We evaluated our method on multiple benchmark datasets for object detection under adverse weather conditions. The model was trained solely on VOC-FOG and VOC-NORMAL, without utilizing training or validation images from RTTS, FOG-Driving, or Foggy Cityscapes. These three datasets were exclusively employed for cross-domain evaluation to validate the model's generalization capability across real-world foggy conditions and synthetic fog scenarios. The dataset sizes are summarized in Table I.

a) VOC-FOG: A synthetic fog dataset generated from PASCAL VOC using an atmospheric scattering model. We use 9,578 foggy images for training, unified into five traffic-related categories.

b) VOC-NORMAL: The clear-image counterpart of VOC-FOG, consisting of 9,578 images from PASCAL VOC (2007+2012), with the same five categories.

c) RTTS: A real-world foggy subset from RESIDE containing 4,322 annotated traffic surveillance images. It is mainly used to test robustness in real hazy environments.

d) FOG-Driving: A real driving dataset collected under different fog densities, providing annotations for common traffic participants. It is used to evaluate performance under real low-visibility conditions.

e) Foggy Cityscapes: A synthetic fog version of Cityscapes generated with a physical fog model. We use a refined subset of 498 images for evaluation in synthetic urban fog scenarios.

Overall, the dataset is sufficiently comprehensive to support evaluation under diverse conditions, thereby enabling a thorough validation of DW-YOLO's detection performance in adverse weather.

B. Experimental Details

DW-YOLO is constructed by inserting the proposed DCAM into the C2f modules at the P3 and P4 multi-scale feature levels. The regression branch adopts the WIoU loss as the bounding box regression objective to enhance localisation

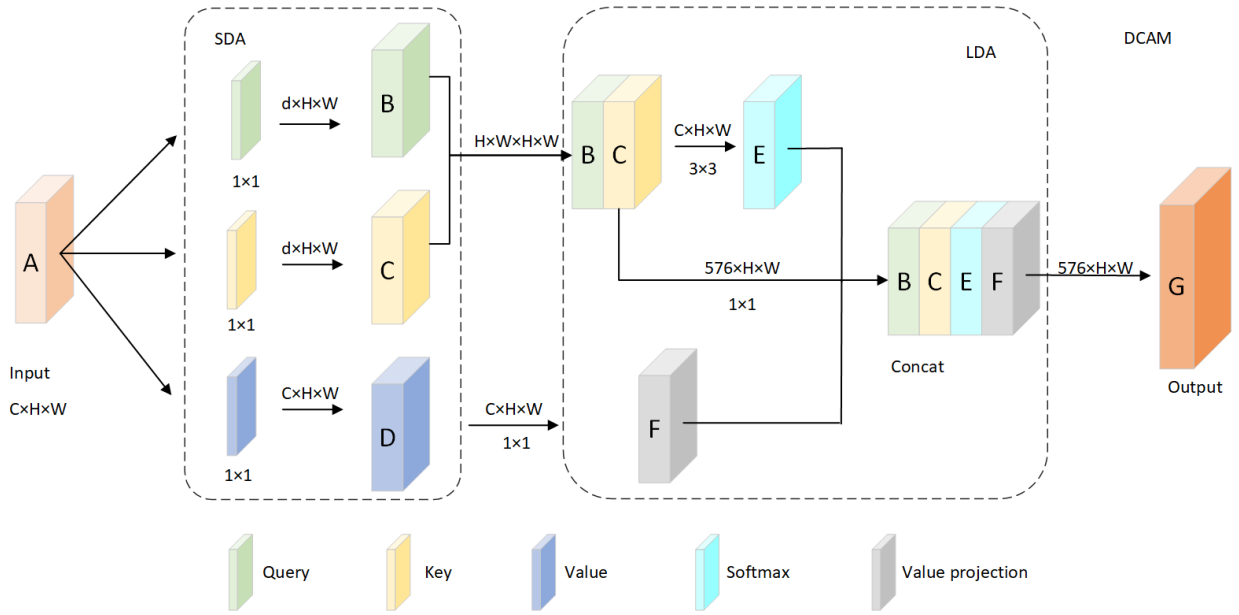


Fig. 2. Structure of the proposed Dual Context Attention Module (DCAM), which integrates Short-Distance Attention (SDA) for local geometric feature modeling and Long-Distance Attention (LDA) for global contextual dependency learning. The module employs multi-branch feature extraction, attention-based similarity computation, and dynamic positional bias to enhance feature representation under complex weather conditions.

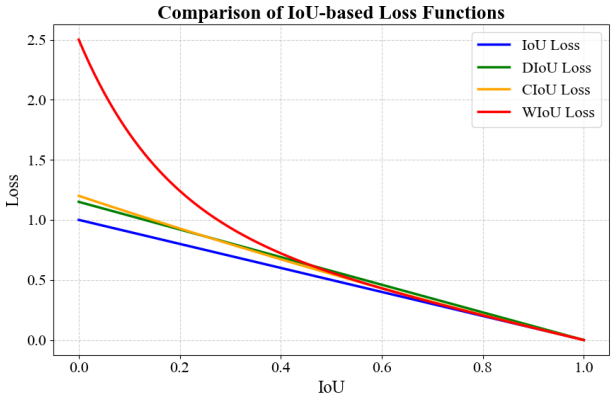


Fig. 3. Comparison of different loss functions during training.

TABLE I
THE NUMBER OF IMAGES IN EACH DATASET AND THE NUMBER OF ANNOTATED OBJECTS

Dataset	Images	Person	Bicycle	Car	Motor	Bus
VOC–FOG [17]	9578	15288	688	1944	663	539
VOC–NORMAL [17]	9578	15288	688	1944	663	539
RTTS [17]	4322	7950	534	18415	862	1838
FOG–Driving [16]	101	269	17	425	7	17
Foggy Cityscapes [16]	498	1715	525	5291	100	43

robustness under low-contrast and occlusion scenarios.All models were trained with an input resolution of 640×640 for 100 epochs using a batch size of 16. The optimizer followed YOLOv8's automatic selection strategy and ultimately employed stochastic gradient descent with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5×10^{-4}. A cosine annealing learning rate scheduler with a short warm-up phase was applied. Automatic mixed precision training was enabled to accelerate convergence and reduce memory consumption. The training data consisted of a joint training set formed by combining VOC-FOG and VOC-NORMAL samples with a 1:1 mixing ratio. All datasets were unified into five object categories: *person*, *car*, *bus*, *bicycle*, and *motorbike*. To improve generalisation under domain shifts, various data augmentation techniques and geometric perturbations were applied during training. In addition, the 1×1 convolutional weights of the Distribution Focal Loss (DFL) regression layer in the detection head were frozen throughout training.All experiments were conducted on a single NVIDIA RTX 3090 GPU.

C. Evaluation Indicators

To comprehensively evaluate the detection performance of the proposed method, we adopt standard metrics from the COCO evaluation protocol, including mAP@0.5:0.95, mAP@0.5, as well as the inference efficiency metric FPS (Frames Per Second). Given a set of predicted bounding boxes B_p and ground truth bounding boxes B_{gt}, the Intersection over Union (IoU) is defined as:

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}, \tag{8}$$

where $|\cdot|$ denotes the area of a region.

For a single category, precision and recall are computed as:

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \qquad \mathrm{Recall} = \frac{TP}{TP + FN}, \tag{9}$$

where TP, FP, and FN represent true positives, false positives, and false negatives.

Based on these, the Average Precision (AP) is defined as:

$$AP = \int_0^1 P(R)\, dR, \tag{10}$$

where $P(R)$ denotes precision at different recall levels.

Following the COCO protocol, two main accuracy indicators are adopted:

mAP@0.5:0.95: Mean Average Precision averaged over IoU thresholds from 0.50 to 0.95 with a step of 0.05:

$$\mathrm{mAP}_{50:95} = \frac{1}{10} \sum_{t=0.50}^{0.95} AP^{\mathrm{IoU}=t}. \tag{11}$$

mAP@0.5: Mean Average Precision at IoU = 0.50:

$$\mathrm{mAP}_{50} = AP^{\mathrm{IoU}=0.50}. \tag{12}$$

In addition to accuracy-oriented metrics, we further evaluate the inference efficiency:

FPS (Frames Per Second): FPS measures the number of images processed per second during inference. Following the timing decomposition of existing real-time detection frameworks, FPS can be expressed as:

$$\mathrm{FPS} = \frac{1000}{T_{\mathrm{pre}} + T_{\mathrm{infer}} + T_{\mathrm{post}}}, \tag{13}$$

where T_{pre}, T_{infer}, and T_{post} denote the preprocessing, inference, and postprocessing time per image (in milliseconds).

Overall, mAP@0.5:0.95 reflects robustness under strict IoU thresholds, mAP@0.5 measures performance under a lenient IoU threshold, and FPS indicates real-time execution capability. Combined, these three metrics provide a comprehensive evaluation of both detection accuracy and runtime efficiency.

D. Comparisons With State-of-the-Arts

We evaluate DW-YOLO on three adverse-weather benchmarks: FOG-Driving, RTTS, and Foggy Cityscapes.

a) FOG-Driving Dataset: Table II and Table III present the results on the FOG-Driving dataset. As shown in Table II, DW-YOLO achieves an mAP@0.5:0.95 of 0.397, outperforming all compared methods and improving over D-YOLO [26] from 0.335 to 0.397. Notably, larger gains are observed on challenging categories such as *bicycle* and *motor*, indicating improved robustness under low-visibility and occlusion conditions.

The mAP@0.5 results in Table III further confirm the effectiveness of our method. DW-YOLO achieves 0.488 mAP@0.5, surpassing YOLOv7 [4], YOLOv10 [6], and TTSDA-YOLO [27]. These improvements can be attributed to the dual-context attention and WIoU-based optimization, which enhance feature representation and improve localization stability in foggy scenes.

TABLE II
COMPARATIVE RESULTS ON THE FOG-DRIVING DATASET (MAP@0.5:0.95)

Method	Person	Bicycle	Car	Motor	Bus	All
YOLOv8 [5]	0.236	0.371	0.450	0.063	0.422	0.308
AOD-YOLOv8 [19]	0.283	0.185	0.543	0.093	0.392	0.299
MSBDN-YOLOv8 [20]	0.225	0.203	0.511	0.127	0.334	0.280
GridDehaze [18]	0.256	0.211	0.498	0.114	0.369	0.290
DCP-YOLOv8 [7]	0.200	0.160	0.500	0.146	0.345	0.270
IA-YOLO [21]	0.227	0.174	0.506	0.076	0.355	0.268
DSNet [35]	0.235	0.155	0.500	0.051	0.378	0.264
MS-DA YOLO [14]	0.241	0.186	0.473	0.083	0.349	0.266
D-YOLO [26]	**0.307**	0.227	**0.576**	0.151	0.412	0.335
DW-YOLO	0.292	**0.529**	0.492	**0.229**	**0.441**	**0.397**

TABLE III
COMPARATIVE RESULTS ON THE FOG-DRIVING DATASET (MAP@0.5)

Method	Person	Bicycle	Car	Motor	Bus	All
YOLOv8 [5]	0.308	0.418	0.589	0.0063	0.435	0.351
YOLOv5 [5]	0.312	0.206	0.574	0.0000	0.440	0.306
YOLOv10 [6]	0.226	0.237	0.545	0.4860	0.332	0.278
YOLOv7 [4]	0.338	0.474	0.624	0.1450	0.496	0.416
FFA+YOLOv7 [36]	0.335	0.473	0.623	0.1450	0.499	0.415
AOD-YOLOv7 [19]	0.324	0.478	0.605	0.0720	0.420	0.380
MSBDN-YOLOv7 [20]	0.263	0.348	0.584	0.0480	0.474	0.344
R-YOLO [25]	0.314	0.352	0.618	0.1470	0.458	0.378
TTSDA-YOLO [27]	0.369	0.467	0.643	0.1450	0.475	0.422
DW-YOLO	**0.436**	**0.529**	**0.666**	**0.2860**	**0.525**	**0.488**

b) RTTS Dataset: Table IV presents the results on the RTTS dataset. DW-YOLO achieves the highest mAP@0.5:0.95 of 0.469, outperforming both dehazing-based methods (e.g., AOD-YOLOv8 [19], GRIDDEHAZE [18]) and recent detectors such as D-YOLO [26]. Notably, larger gains are observed for *motor* and *bus*, which are more affected by scale variation and motion blur in adverse conditions. These results demonstrate that the proposed method generalizes well to real-world foggy scenes without requiring explicit image restoration, benefiting from the dual-context attention and WIoU-based optimization.

c) Foggy Cityscapes Dataset: Table V summarizes the results on the Foggy Cityscapes-refined dataset. DW-YOLO achieves an mAP@0.5:0.95 of 0.357, outperforming all compared methods. Compared with YOLOv8, our method significantly improves detection performance, and it also surpasses D-YOLO [26] across all categories. Notably, larger gains are observed for small and low-contrast objects such as *bicycle* and *motor*, indicating enhanced robustness under challenging conditions. These results demonstrate strong cross-domain generalization from synthetic to real foggy scenes, benefiting from the proposed dual-context attention and WIoU-based optimization.

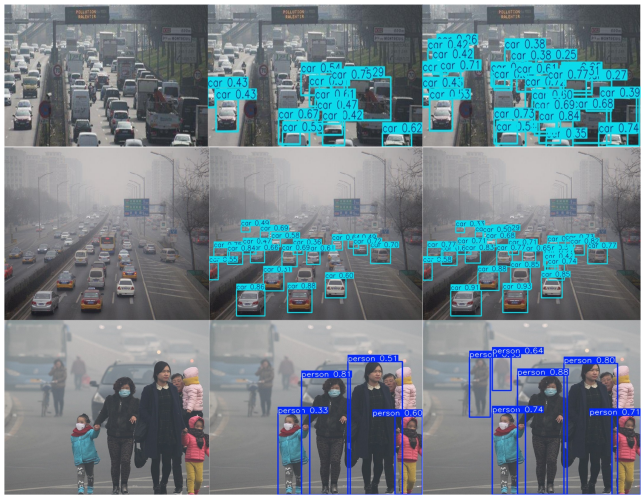
Overall, the consistent improvements across datasets demonstrate that DW-YOLO effectively mitigates domain shifts and enhances feature robustness under adverse weather conditions.

E. Qualitative Results

To further validate the effectiveness of the proposed method, we provide qualitative comparisons on both real-world and synthetic foggy scenes, as shown in Fig. 4. Compared with

TABLE IV
COMPARATIVE RESULTS ON THE RTTS DATASET (MAP@0.5:0.95)

Method	Person	Bicycle	Car	Motor	Bus	All
YOLOv8 [5]	0.578	0.331	0.450	0.253	0.275	0.377
AOD-YOLOv8 [19]	0.598	0.358	0.407	0.233	0.130	0.345
MSBDN-YOLOv8 [20]	0.589	0.374	0.393	0.209	0.120	0.337
GRIDDEHAZE [18]	0.612	0.386	0.453	0.258	0.146	0.371
DCP-YOLOv8 [7]	0.621	0.393	0.417	0.237	0.139	0.361
IA-YOLO [21]	**0.671**	0.353	0.414	0.211	0.136	0.357
DSNET [35]	0.566	0.345	0.402	0.198	0.124	0.327
MS-DAYOLO [14]	0.637	0.391	0.479	0.281	0.157	0.389
D-YOLO [26]	0.658	0.402	**0.538**	0.308	0.242	0.430
DW-YOLO	0.568	**0.467**	0.517	**0.396**	**0.398**	**0.469**



(a) Ground Truth (b) YOLOv8 (c) Ours

Fig. 4. Detection results of our proposed method. (a) Ground truth, (b) YOLOv8, (c) our DW-YOLO. As we can see, DW-YOLO produces more complete and accurate bounding boxes compared with YOLOv8.

YOLOv8, DW-YOLO yields clearer and more stable detections under dense fog. It effectively reduces missed detections and false positives, produces more accurate bounding boxes for distant and low-contrast objects, and demonstrates clear advantages on small targets such as pedestrians. These visualization results further confirm the robustness and reliability of DW-YOLO in complex foggy environments.

F. Ablation Study

To further verify the effectiveness of each component in DW-YOLO, we conduct an ablation study on the FOG-Driving dataset. The variants V0–V6 in Table VI are constructed by gradually activating three key modules on top of the YOLOv8 baseline, including the proposed Dual-Context Attention Module (DCAM), the WIoU-based regression loss, and the data augmentation (AUG) strategies tailored for adverse weather.

V0 corresponds to the vanilla YOLOv8 detector, which attains an mAP@0.5:0.95 of 0.308 on FOG-Driving and serves as our reference. Introducing DCAM (V1–V2) consistently improves detection accuracy, reflecting the benefit of leveraging dual-context features for robust representation learning under fog-induced visibility degradation. When WIoU is further employed as the bounding box regression loss (V3), the localization quality is enhanced, yielding additional gains over the DCAM-only variants.

Applying weather-aware data augmentation (V4) brings further improvements by exposing the model to diverse fog densities and geometric perturbations during training. Finally, the full configuration that jointly integrates DCAM, WIoU, and AUG (V6) achieves the best performance, reaching an mAP@0.5:0.95 of 0.397 on FOG-Driving. This represents a clear improvement of 8.9 percentage points over the baseline of YOLOv8, confirming that the proposed components are complementary and jointly contribute to robust object detection in adverse weather conditions.

TABLE V
COMPARATIVE RESULTS ON THE FOGGY CITYSCAPES DATASET (MAP@0.5:0.95)

Method	Person	Bicycle	Car	Motor	Bus	All
YOLOv8 [5]	0.157	0.087	0.318	0.058	0.124	0.149
AOD-YOLOv8 [19]	0.267	0.174	0.373	0.233	0.192	0.210
MSBDN-YOLOv8 [20]	0.251	0.143	0.369	0.209	0.167	0.195
GRIDDEHAZE [18]	0.274	0.212	0.353	0.258	0.179	0.210
DCP-YOLOv8 [7]	0.236	0.144	0.332	0.237	0.129	0.174
IA-YOLO [21]	0.235	0.127	0.341	0.211	0.130	0.177
DSNET [35]	0.221	0.137	0.323	0.198	0.133	0.169
MS-DAYOLO [14]	0.243	0.141	0.339	0.281	0.161	0.183
D-YOLO [26]	0.311	0.204	0.420	0.308	0.215	0.245
DW-YOLO	**0.320**	**0.379**	**0.497**	**0.396**	**0.252**	**0.357**

TABLE VI
ABLATION STUDY

Module	V0	V1	V2	V3	V4	V5	V6
YOLOv8	✓	✓	✓	✓	✓	✓	✓
DCAM		✓			✓	✓	✓
WIoU				✓	✓		✓
AUG			✓			✓	✓
FOG–Driving	0.308	0.352	0.361	0.357	0.369	0.390	**0.397**

G. Efficiency Analysis

We evaluate our DW-YOLO in efficiency.Experiments were conducted on the RTTS dataset using the mAP@0.5:0.95 metric, with all tests performed on a single RTX 3090 GPU. The input image resolution was set to 640×640×3. We compared our model against various defogging and detection methods. As shown in Fig. 5, our model demonstrates clear advantages in terms of parameter count and inference speed, ensuring real-time prediction performance.

V. CONCLUSION

In this work, we presented DW-YOLO, a robust object detection framework designed for adverse weather and cross-domain scenarios. The proposed method integrates dual-context feature modeling, WIoU-based regression, and weather-aware augmentation into a unified architecture. Experiments on RTTS, FOG-Driving, and Foggy Cityscapes demonstrate consistent performance gains over representative baseline detectors.

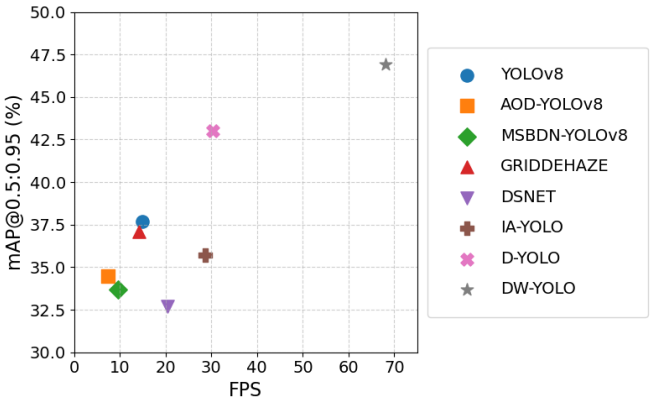


Fig. 5. Inference speed comparison (FPS) and mAP@0.5:0.95 of different models.

In addition to accuracy improvements, DW-YOLO achieves higher inference speed through lightweight architectural design, yielding noticeably higher FPS than most existing robust detection approaches. This makes the framework more suitable for real-time deployment in safety-critical applications.

Overall, DW-YOLO effectively mitigates the performance degradation caused by adverse weather and domain shifts, providing a practical solution for perception in degraded environments. Future work will focus on further improving adaptability and deployment efficiency, including lighter network designs and potential integration with multi-modal sensing.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[4] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.

[5] G. Jocher, "Yolov5 and yolov8," https://github.com/ultralytics, 2023.

[6] A. Wang, H. Chen, L. Liu *et al.*, "Yolov10: Real-time end-to-end object detection," in *Advances in Neural Information Processing Systems*, 2024.

[7] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[8] W. Ren, S. Liu, H. Zhang *et al.*, "Single image dehazing via multi-scale convolutional neural networks," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 154–169.

[9] T.-Y. Lin, P. Dollár, R. Girshick *et al.*, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[10] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[12] Y. Chen, W. Li, C. Sakaridis *et al.*, "Domain adaptive faster r-cnn for object detection in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] S. Zhang, H. Tuo, J. Hu *et al.*, "Domain adaptive yolo for one-stage cross-domain detection," in *Asian Conference on Machine Learning*, 2021.

[14] M. Hnewa and H. Radha, "Multiscale domain adaptive yolo for cross-domain object detection," in *IEEE International Conference on Image Processing*, 2021.

[15] Z. Tong, Y. Chen, Z. Xu *et al.*, "Wise-iou: Bounding box regression loss with dynamic focusing mechanism," *arXiv preprint arXiv:2301.10051*, 2023.

[16] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[17] B. Li, W. Ren, D. Fu *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2019.

[18] X. Liu, Y. Ma, Z. Shi *et al.*, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7314–7323.

[19] B. Li, X. Peng, Z. Wang *et al.*, "Aod-net: All-in-one dehazing network," in *IEEE Conference on Computer Vision (ICCV)*, 2017.

[20] H. Dong, J. Pan, L. Xiang *et al.*, "Multi-scale boosted dehazing network with dense feature fusion," in *CVPR*, 2020.

[21] W. Liu, G. Ren, R. Yu *et al.*, "Image-adaptive yolo for object detection in adverse weather conditions," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 1792–1800.

[22] Y. Ogino, Y. Shoji, T. Toizumi *et al.*, "Erup-yolo: Enhancing object detection robustness for adverse weather condition by unified image-adaptive processing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[23] Z. Liu, T. Fang, H. Lu *et al.*, "Masfnet: Multiscale adaptive sampling fusion network for object detection in adverse weather," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[24] Y. Yuan, Y. Wei, Y. Guo *et al.*, "Cme-yolo: A cross-modal enhanced yolo algorithm for adverse weather object detection in autonomous driving," *Big Data and Cognitive Computing*, vol. 9, no. 4, 2025.

[25] L. Wang, H. Qin, X. Zhou *et al.*, "R-yolo: A robust object detector in adverse weather," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[26] Z. Chu, "D-yolo: A robust framework for object detection in adverse weather conditions," *arXiv preprint arXiv:2403.09233*, 2024.

[27] M. Zhang, Q. Rong, H. Jing *et al.*, "Ttsda-yolo: A two training stage domain adaptation framework for object detection in adverse weather," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[28] N. Carion, F. Massa, G. Synnaeve *et al.*, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[29] Z. Fu, K. Chang, M. Ling *et al.*, "Auxiliary domain-guided adaptive detection in adverse weather conditions," in *Asian Conference on Computer Vision*, 2024.

[30] R. Zhao, H. Yan, and S. Wang, "Revisiting domain-adaptive object detection in adverse weather," in *European Conference on Computer Vision*, 2024.

[31] B. Li, J. Li, X. Liu *et al.*, "V2x-dgw: Domain generalization for multi-agent perception under adverse weather conditions," in *IEEE International Conference on Robotics and Automation*, 2025.

[32] Z. Zheng, P. Wang, W. Liu *et al.*, "Distance-iou loss: Faster and better learning for bounding box regression," in *AAAI Conference on Artificial Intelligence*, 2020.

[33] Z. Zheng, P. Wang, D. Ren *et al.*, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8574–8586, 2021.

[34] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[35] S. Huang, T. Le, and D. Jaw, "Dsnet: Joint semantic learning for object detection in inclement weather conditions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2623–2633, 2020.

[36] X. Qin, Z. Wang, Y. Bai *et al.*, "Ffa-net: Feature fusion attention network for single image dehazing," in *AAAI*, 2020.