SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I6P107

Volume 13 Issue 6, 78-97, June 2026

Original Article

DATNet-RS: Domain-Adaptive Temporal Attention with Residual Shrinkage and Online PSO for Robust Object Detection in Adverse Weather

Joseph Rish Simenthy[1], Pallavi Singh[2]

[1,2]*Department of Electronics and Communication Engineering, Hindustan Institute of Technology and Science, Padur, Tamil Nadu, India.*

[1]*Corresponding Author : rp.22703065@student.hindustanuniv.ac.in*

Received: 09 March 2026 Revised: 08 April 2026 Accepted: 07 May 2026 Published: 27 June 2026

Abstract *- Extreme weather conditions are a serious setback to the reliability of object detection in real-world driving conditions. This paper introduces DATNet-RS, a domain adaptive detection system that runs effectively in these conditions and does not require retraining or labelled weather data. A six-component multi-scale attention module is proposed, including local and global versions of channel, spatial, and temporal attention, which, in combination, reduce the noise-enhancing feature channels, highlight spatially consistent object regions, and capitalize on the frame-to-frame temporal continuity. Second, there is a common residual shrinkage denoising block on all levels of the feature pyramid to reduce the low-amplitude noise activations caused by weather, but not the structurally informative responses. Third, a gradient-free online inference-time adaptation scheme is added, where a small nine-dimensional parameter vector - controlling all attention magnitudes and shrinkage thresholds - is jointly optimized by the use of Particle Swarm Optimization (PSO). Experiments demonstrate that DATNet-RS continuously improves the performance compared to the baseline in all four adverse conditions and increases average mAP@0.5 to 75.1% (+4.3 percentage points) and average mAP@0.5:0.95 to 44.2% (+2.9 percentage points), in addition to maintaining real-time performance of GPU inference at about 50 FPS. On mAP at 0.5, the improvement of per-condition is between +3.9% (night) to +4.9% (rain). The assessment is done by isolating the contribution of each component in a seven-configuration ablation study and making comparisons between these and YOLOv9, RT-DETR, and Deformable DETR position DATNet-RS, a highly competitive real-time detector in the evaluated set.*

Keywords *- Object detection, Attention mechanisms, Temporal modeling, Residual Shrinkage, Particle Swarm Optimization, Feature Pyramid Network.*

1. Introduction

One of the main perceptual tasks of intelligent transportation and autonomous driving is object detection [1, 2], which allows cars to recognize and position pedestrians, cyclists, motorcycles, cars, buses, trucks, and road infrastructure in real time. Significant advancements have been made based on deep convolutional architectures [3-5], an anchor-free paradigm [6, 7], and large-scale annotated benchmarks [8, 9]. However, a persistent limitation exists: sensors that are trained in bright conditions often exhibit drastic deterioration of their performance in challenging situations, which is typical of the on-road conditions and involves safety issues directly [10, 11].

Every unfavourable circumstance creates a physically unique type of loss. Fog will weaken the light in both aerosol scattering and attenuate the contrast of the scene, and will also obliterate the boundaries of objects, especially those at a range of more than 30-50 meters [12]. The rain also causes spatially anisotropic streaking and dynamically changing noise patterns in different directions depending upon the direction of the wind and the strength of the precipitation [13]. Snow overlays introduce almost white occlusions on the scene elements, decreasing the distinction in texture between objects and the background, and generating high miss detection rates of spatially compact objects [14]. The degradation of the nighttime signal-to-noise ratio across all the channels is highly degraded, artificial sources of high-dynamic-range problems are introduced with dramatic severity, and intermittent deep-shadow areas are created, wholly covering object signatures [15]. Each phenomenon introduces a distributional shift in comparison to the statistics of features seen during training, which are expressed as high false-negative rates, low confidence, and localization inconsistency [16]. The adverse weather robustness has been previously covered under three general strategies. Image restoration algorithms seek to pre-

process inputs to degrade them out prior to conventional detection [17, 18]. Domain adaptation methods such as adversarial alignment [19], distribution normalization [20], and style transfer [21] build bridges between the source and target gaps during training. The data augmentation methods generate synthetic weather effects to increase the train distribution [22, 23]. Although both strategies offer valuable benefits, they have a major restriction in common: they are all fixed in behaviour once the model has been trained. A test-time condition that lies outside the training distribution may lead to poorer performance that occasionally cannot be predicted [24, 25]. Conditions may change rapidly within one sequence, thus grossly decreasing performance. The role images can play in the defining complementary direction to which more and more attention is devoted is the so-called test-time or online adaptation, where model parameters or predictions are adjusted during inference according to the current input statistics [26-28]. Gradient-based test-time adaptation [26, 27] is effective but very expensive and prone to instability unless monitored. Within the latency bound, given the lightweight derivative-free variants that are trying to optimize a restricted parameter set, is a more realistic direction. It encourages a hybrid strategy: a highly robust architecture in terms of structural robustness due to specific inductive biases, coupled with a very efficient inference-time adaptation mechanism responding to condition changes that are observed.

Recent studies between 2023 and 2026 have increasingly explored lightweight transformer detectors [40], adaptive perception systems [60], robust video understanding [61], and inference-time adaptation for autonomous driving under environmental uncertainty [62, 63]. Modern detectors such as YOLOv10 [64], YOLOv11 [65], YOLOv12 [76], and recent adaptive perception frameworks [66] have demonstrated improved robustness and efficiency; however, most existing approaches still rely on fixed post-training behaviour or computationally intensive adaptation strategies. This highlights the continuing need for lightweight and dynamically adaptive detection systems suitable for real-time deployment under rapidly changing adverse-weather conditions.

This paper presents Domain-Adaptive Temporal attention Network with Residual Shrinkage (DATNet-RS). The DATNet-RS enhances strength in two mutually supportive tiers. At the architectural level, it adds a six-component multi-scale attention module to a lightweight YOLO-type detector and a shared block of residual shrinkage denoising block, which complement each other in the quality of features in poor inputs. On the deployment level, it presents a gradient-free and label-free online adaptation algorithm that employs Particle Swarm Optimization to modify a small nine-dimensional parameter set - determining attention values and shrinkage thresholds - according to specified changes in the scene itself. This two-level model deals with the technical drawbacks of fixed models, as well as the unrealism of costly online optimization.

DATNet-RS is introduced as a domain-adaptive detection framework for robust object detection in inclement weather scenarios. At the architectural level, it augments a lightweight YOLO-style detector with a six-component multi-scale attention module and a shared residual shrinkage denoising block to strengthen informative object features while suppressing weather-induced corruption. At the deployment level, it incorporates a gradient-free, label-free online adaptation mechanism based on Particle Swarm Optimization (PSO), which adjusts a compact nine-dimensional parameter vector controlling attention strengths and shrinkage thresholds according to scene changes observed during inference. This framework addresses a critical gap in existing detection systems, in which model behaviour remains fixed after training and cannot adapt effectively to dynamic deployment conditions. Unlike approaches that depend on either static robustness design or computationally expensive gradient-based test-time adaptation, DATNet-RS provides a lightweight, stable, and real-time compatible solution for inference-time adaptation.

The principal contributions of this work are:

- Multi-Scale Six-Component Attention Module: It suggests a new attention subsystem to be able to model both local and global dependencies between channel, spatial, and temporal features. This enables the noise-amplified channels to be suppressed by the network, object regions to be concentrated on, and temporal frame consistency to be used, three different but complementary approaches to adverse-weather robustness.
- Shared Residual Shrinkage Denoising: A shared-weight residual shrinkage block with learner-bounded thresholds is used across all three feature pyramid levels, directly aimed at suppressing low-amplitude noise-like activations without changing training data or augmenting data pipes and with minimal additional parameter cost.
- Online Gradient-Free PSO Adaptation: It proposes an environment-sensitive mechanism enabling the PSO-based optimization of a nine-dimensional parameter vector every time a change in the environment of interest is suggested by scene statistics. It is label-free and backpropagation-free, and can be run on real-time latency budgets
- Standardized Evaluation and Ablation: Multi-run statistics of mAP@0.5 and mAP@0.5:0.95 are reported in experiments of ACDC [29]. A seven-configuration ablation test separates the contribution of each element, and a breakdown of latency validation of computational viability.

The balance of this research study has the following structure. Section 2 represents an overview of related works. In Section 3, the DATNet-RS architecture will be explained.

Experimental details are presented in Section 4. Results for the experiments are described in Section 5. Ablation studies are outlined in Section 6. An explanation of deployment challenges, potential failures, and future work possibilities is provided in Section 7. The conclusion of the paper can be found in Section 8.

Fig. 1 Representative ACDC samples across adverse conditions

2. Related Work

2.1. Object Detection: Foundations and Real-Time Architectures

Two-stage object-detection models typically consist of either a two-stage model (as found in R-CNN [3], as well as enhanced versions such as Fast R-CNN [30] & Faster R-CNN [4]) or a one-stage model that first generates region-proposals and then classifies and regresses on each proposal. The limitations of using two-stage models are that they can be too computationally complex to use in "real-time" applications. Single detectors reduce this limitation since they eliminate the need for generating region proposals and perform all of the detection within a single pass. YOLO [1] introduced a unified regression-based framework enabling real-time performance, with subsequent improvements in YOLOv2 [31], YOLOv3 [32], YOLOv4 [33], YOLOv5 [34], YOLOv7 [35], and YOLOv9 [36]. Anchor-free methods such as FCOS [6], CenterNet [7], and ATSS [37] further simplify detection by removing anchor design constraints. Transformer-based detectors, including DETR [38], Deformable DETR [39], and RT-DETR [40], model global dependencies effectively but often incur higher computational cost.

Despite these advancements, most detection architectures are optimized for standard imaging conditions and exhibit limited robustness under adverse weather, as they do not explicitly address noise corruption, domain shifts, or temporal instability. Transformer-based models, in particular, may face latency challenges in real-time deployment. In contrast, the proposed framework adopts a lightweight CNN-based design using MobileNetV3 [41] and FPN [42], with an emphasis on both robustness and real-time performance. Recent detector developments, including YOLOv10 [64], YOLOv11 [65], YOLOv12 [86], and efficient hybrid transformer-CNN architectures [40], have further emphasized end-to-end real-time detection and deployment-oriented optimization for autonomous perception systems, particularly in edge-constrained environments [67, 68, 81].

2.2. Adverse Weather Detection

Adverse-weather object detection has gained increasing attention with the introduction of dedicated benchmarks such as RTTS, Foggy Cityscapes [24], and ACDC [29], which provide realistic evaluation scenarios. Image restoration methods, including dehazing [45], deraining [13], and low-light enhancement [17], aim to improve visual quality prior to detection. However, such pre-processing steps may introduce artifacts that are not aligned with detector training distributions, thereby limiting performance [24]. Domain adaptation approaches, including adversarial alignment [19], distribution normalization [20], and curriculum learning, attempt to reduce domain shifts but typically require access to target-domain data during training. While these approaches improve robustness under controlled conditions, their behaviour remains largely fixed after training. As a result, their effectiveness is reduced when encountering unseen or

rapidly changing environmental conditions during deployment. The proposed framework addresses this limitation by enabling adaptation during inference without requiring retraining. More recent studies have explored weather-robust perception using uncertainty-aware learning, multi-condition adaptation [69], and lightweight enhancement-guided detection pipelines [70]; however, many approaches still depend on offline retraining or computationally expensive restoration stages, underscoring the need for efficient online adaptive solutions. Very recent studies, including adverse-weather YOLO variants evaluated on ACDC [77], [78], domain-adaptive detection methods addressing foggy and rainy driving [80], and multi-condition perception frameworks [86], further confirm the ongoing relevance of robust detection under realistic driving degradation, yet none incorporate inference-time gradient-free adaptation as proposed in this work.

2.3. Attention Mechanisms in Feature Learning

Attention mechanisms improve feature representation by emphasizing relevant information and suppressing noise. Squeeze-and-Excitation Networks (SENet) [46] introduced channel-wise recalibration, while CBAM [47] extended this concept to sequential channel and spatial attention. Non-local networks [48] capture long-range dependencies through self-attention, although at increased computational cost. Coordinate Attention [49] incorporates positional information into channel attention. In adverse-weather scenarios, attention mechanisms help suppress noise-affected feature channels and focus on stable object regions [50]. Temporal attention further leverages frame-to-frame consistency in video sequences [51], [52]. However, most existing approaches treat channel, spatial, and temporal attention independently, limiting their ability to jointly address complex degradation patterns. In contrast, the proposed framework integrates channel, spatial, and temporal attention in both local and global forms within a unified multi-scale design, enabling a more comprehensive and coordinated response to adverse-weather effects. Recent lightweight attention architectures have also focused on balancing global contextual modelling with deployment efficiency, particularly for real-time video understanding and autonomous perception applications [61, 71, 72].

2.4. Feature-Space Denoising and Residual Shrinkage

Feature-space denoising operates on intermediate representations to suppress noise while preserving useful information. Residual shrinkage methods, such as Deep Residual Shrinkage Networks (DRSN) [54], apply soft-thresholding to eliminate low-amplitude activations while retaining significant features. These approaches have demonstrated effectiveness in signal processing and fault diagnosis tasks. However, their application to object detection under adverse weather remains limited, particularly in the context of multi-scale feature representations where consistent denoising is required across different resolutions. The proposed framework extends residual shrinkage into a shared multi-scale design, enabling efficient and consistent noise suppression across all feature pyramid levels.

2.5. Test-Time and Online Adaptation

Test-Time Adaptation (TTA) enables models to adjust their behaviour during inference using unlabelled data. Methods such as TTT [26], Tent [27], and MEMO [28] rely on gradient-based optimization to update model parameters, improving robustness under distribution shifts. However, these approaches introduce additional computational overhead and may affect stability, making them less suitable for real-time systems. Derivative-free optimization methods, including evolutionary strategies and Particle Swarm Optimization (PSO) [56], offer an alternative by optimizing parameters without requiring gradients. Nevertheless, existing approaches either rely on expensive backpropagation or lack structured integration with detection architectures. In contrast, the proposed framework employs a lightweight PSO-based adaptation strategy operating on a compact parameter space, enabling stable and efficient inference-time adaptation under real-time constraints. Recent inference-time adaptation research has increasingly focused on lightweight parameter-efficient adaptation [62], entropy-aware optimization [63], and stable online adaptation for safety-critical systems [73]. Continual test-time adaptation methods [74] and source-free domain adaptation frameworks [75] have demonstrated promising results; nevertheless, many current methods still rely on gradient updates, auxiliary memory modules, or computationally intensive optimization procedures that limit practical deployment in real-time autonomous driving scenarios. More recent surveys [66] and empirical studies [77] have reinforced this observation, noting that the gap between laboratory TTA methods and real-time deployable solutions remains substantial as of 2025.

2.6. Closing Synthesis

Existing research demonstrates that object detection architectures, attention mechanisms, denoising techniques, and test-time adaptation strategies each contribute to improving robustness under adverse conditions. However, these approaches are typically developed independently, focusing either on architectural enhancement or adaptive behaviour without a unified design. In contrast, the proposed framework integrates multi-scale attention, residual shrinkage denoising, and lightweight gradient-free adaptation within a single detection pipeline. This unified approach enables both structural robustness and dynamic adaptation, allowing the model to respond effectively to changing environmental conditions.

3. Proposed Method

3.1. Baseline Detector

DATNet-RS consists of a lightweight YOLO-style base based on MobileNetV3 [41], which is a hierarchical feature extractor, Feature Pyramid Network (FPN) [42], which is a top-down multi-scale fusion network, and three detection

heads of spatial strides P3, P4, and P5, namely 8, 16, and 32. Given an input,

$$x \in R^{3\times H\times W}$$

The backbone or feature extractor extracts a feature map at three scales and transforms them into a common channel dimension Cd = 256 prior to FPN fusion. Every detection head generates an output tensor of size (A x (5 + C)) per spatial cell, where A = 3 anchors per cell, C is 10 object classes, and the 5 dimensions encode (D x, Dy, D w, D h, objectness). With 7.2M parameters, it takes 45 MB in disk, and at 640x640 resolution, it is 94.5 FPS on an NVIDIA A100 GPU.

The choice of MobileNetV3 was informed by its balanced accuracy and parameter ratio and by bypass characteristics such as depth-wise separable convolutions and hard-swish activations that allow it to be actually deployed on the edge.

In safety-critical embedded systems, dashcams, and forward-facing perception units, reducing the footprint of the static memory without any decrease in the quality of the features is one of the key design goals.

The minimum that is needed is 6.2 GFLOPs per forward pass at 640x640 resolution. DATNet-RS includes a 4.8 GFLOPs attention module (+3.4 GFLOPs), residual shrinkage (+0.2 GFLOPs), and PSO interface overhead (+1.2 GFLOPs), which is a 77% increase over transformer-based detectors like Deformable DETR (86.1 GFLOPs) and RT-DETR (74.2 GFLOPs).

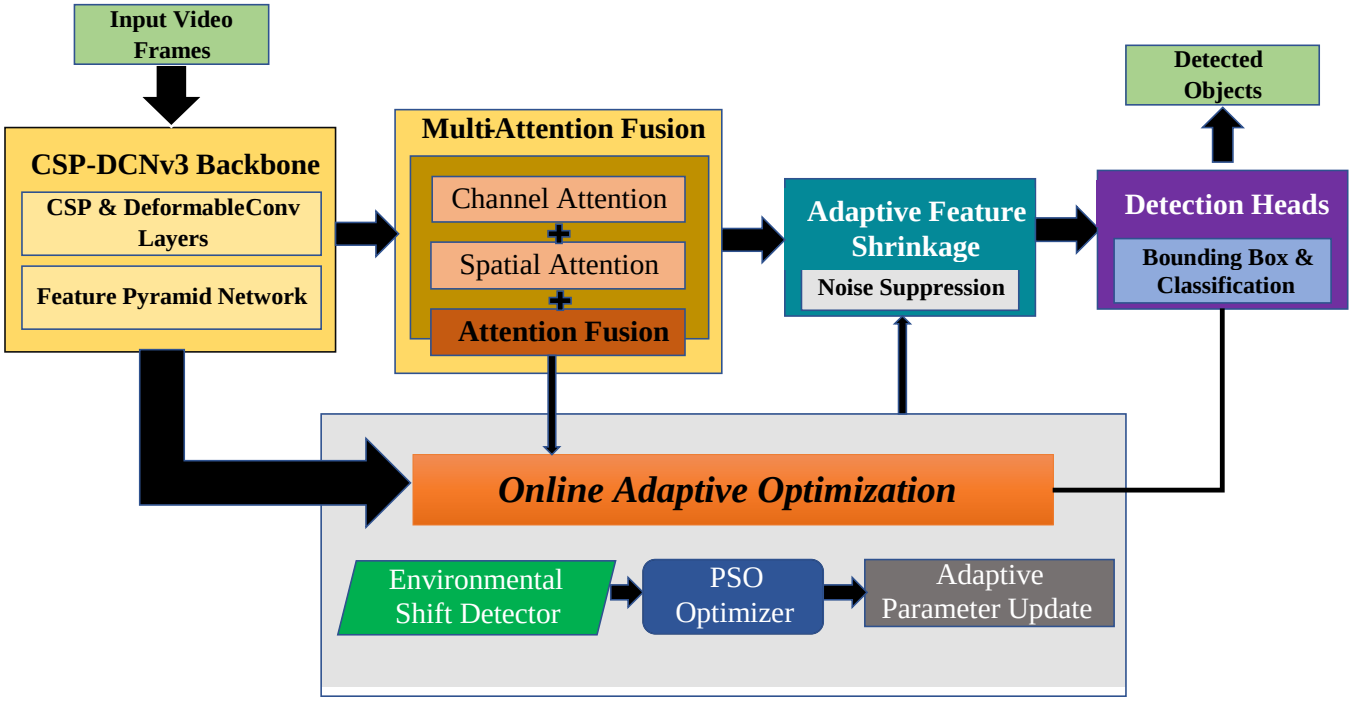

Fig. 2 Overall architecture of DATNet-RS

3.2. DATNet-RS Architecture Overview

The DATNet-RS adds two structural and one deployment-level components to the base: (1) a multi-scale attention module, which is inserted to the output side of the FPN feature fusion, (2) a shared residual shrinkage denoising block, which autonomously operates on the output of each level of the pyramid, and (3) online adaptation, in which one PSO-based adaptation mechanism modulates attention and shrinkage parameters combines with the network during inference.

Backbone and topology of FPN, detection heads of FPN remain the same as in the baseline, meaning that all performance differences are explained by the presence of the three new components and not due to unrelated variations.

3.3. Multi-Scale Six-Component Attention Module

3.3.1. Design Motivation

The feature maps in adverse weather are spoiled in spatial and channel non-uniform manners. Only one method of suppression is not enough: fog causes a reduction in edge-sensitive channels and remote spatial areas in an even manner, whereas rain artifacts are focused and temporal.

The six-component design is complementary, as channel, spatial, and temporal attention cover which aspect dimensions to believe in, where in the spatial map to focus on, and how consistent it is over time.

Local variations are adapted to local patterns; global variations are adapted to large contexts.

3.3.2. Channel Attention

Given feature tensor $F^l \in R^{C\times H\times W}$ At pyramid level l, channel attention computes a per-channel weight vector to recalibrate feature responses. The local channel attention variant applies a 3×3 depthwise convolution to capture inter-channel correlations within a limited receptive field.

The global channel attention variant applies global average pooling to aggregate spatial context into a channel descriptor, followed by a two-layer MLP with reduction ratio r = 16 [46]. The two weight vectors are combined with learnable scalar strengths:

$$F_{ca}^{l} = F^{l}\left(\alpha_1 \cdot w_{lcl}^{l} + \alpha_2 \cdot w_{gbl}^{l}\right)$$

where $w_{lcl}, w_{gbl} \in R^{C}$, $\alpha_1, \alpha_2 \in [0,2]$ are scalar strength parameters forming part of the adaptive parameter vector θ, and denotes channel-wise broadcast multiplication.

3.3.3. Spatial Attention

Spatial attention reweights spatial positions to emphasize object-relevant regions and suppress background clutter or noise patches. Local spatial attention applies a 7×7 convolution over the channel-reduced feature map. Global spatial attention concatenates max-pooled and average-pooled channel summaries and passes them through a 7×7 convolutional layer [47]:

$$F_s a^{l} = F_c a^{l}\left(\alpha^{3} \cdot m_l c l^{l} + \alpha^{4} \cdot m_g b l^{l}\right)$$

where $m_{lcl}, m_{gbl} \in R^{1\times H\times W}$, $\alpha_3, \alpha_4 \in [0,2]$.

3.3.4. Temporal Attention

Temporal attention exploits the structural consistency of object representations across adjacent video frames. Genuine objects produce stable feature patterns while weather artifacts are temporally variable. A short-term memory buffer retains the feature map. $F_{sa}^{(t-1)}$ from the preceding frame (local temporal), while a running exponential moving average

$$EMA^{l}(t) = \beta \cdot EMA^{l}(t-1) + (1-\beta) \cdot F_{sa}^{t}$$

maintains longer-term context (global temporal), with decay β = 0.9. Cosine-similarity gates compute temporal attention weights:

$$F_{ta}^{l} = F_{sa}^{l}\left(\alpha_5 \cdot gate_l(F_{sa}^{t}, F_{sa}^{t-1}) + \alpha_6 \cdot gate_g(F_{sa}^{t}, EMA^{l})\right)$$

where α_5, α_6 [0, 2] are the temporal strength parameters. The gate functions produce spatial weight maps in R^(1×H×W) by computing pixel-wise cosine similarity between the current feature map and the stored state, passed through a sigmoid activation. This formulation upweights regions where features are temporally consistent — indicative of genuine objects — and downweights transient noise regions.

3.3.5. Parameterization for Online Adaptation

The six scalars [α_1, ..., α_6] collectively form the first six dimensions of the adaptive parameter vector θ. Encoding each attention component's strength as a single scalar provides an interpretable and low-dimensional interface between the online optimizer and the network: the optimizer can, for instance, increase temporal attention under rapidly changing noise or increase spatial attention when contrast degrades, without any gradient information.

3.4. Shared Residual Shrinkage Denoising

3.4.1. Motivation and Formulation

Soft-thresholding is a classical signal processing technique for denoising [54]. Given an activation value v, soft-thresholding with threshold τ produces:

$$shrink(v,\tau) = sign(v) \cdot max(|v| - \tau, 0)$$

Activations with magnitude below τ are zeroed; those above are reduced by τ. This selectively eliminates low-amplitude noise-like activations while preserving high-confidence signal responses. Embedded within a residual connection, this operation maintains information flow through the skip path:

$$F_{rs}^{l} = F_{ta}^{l} + \text{shrink}(\text{Conv}(F_{ta}^{l}), \tau)$$

where Conv is a 1×1 projection and the three learnable thresholds τ = [τ_1, τ_2, τ_3] [0, 3] form the remaining three dimensions of θ. The upper bound $\tau \leq 3$ was set empirically to prevent over-suppression of genuine signal activations; values above this threshold were found to remove responses corresponding to real object boundaries in fog scenes.

3.4.2. Shared-Weight Design Across Pyramid Levels

Each of the three pyramid levels, P3, P4, and P5, has a single shrinkage module whose weights are shared. Independent level- by level modules would increase shrinkage parameters with no obvious benefit to quality since weather noise statistics are largely similar over neighboring scales. The common design ensures that there is a uniform denoising prior, the suppression behavior is avoided when the multi-scale is not stable, and parameter overhead is cut to about 45K more parameters. This increases empirically by +1.5% inference latency with an average improvement in mAP of +0.4% over the attention-only setup.

3.5. PSO-Based Online Inference-Time Adaptation

3.5.1. Adaptive Parameter Vector

The nine-dimensional adaptive parameter vector θ jointly encodes all tunable components:

θ = [α_1, α_2, α_3, α_4, α_5, α_6, τ_1, τ_2, τ_3]

with bounds αi [0, 2], i = 1,...,6 and τj [0, 3], j = 1, 2, 3. In the normal inference, then θ is set to its. trained default values. At the scene-change trigger, PSO is looking within these limits to find the configuration that maximizes the outcome of the detection objective in Section 3.5.3. These bounds were

chosen empirically by grid search on the validation set, choosing ranges that allow. non-degenerative architectural modulation meaning meaningful modulation of architecture without permitting degenerate configurations (e.g., zero attention). excessive shrinkage).

3.5.2. Scene-Change Detection Trigger

At each frame t, the DATNet-RS framework computes four scene statistics: mean pixel intensity μ_I, intensity standard deviation σ_I, edge density ρ_E, (defined as the fraction of pixels with local L_1 gradient exceeding a fixed threshold), and Michelson contrast C_MThese statistics are maintained using a sliding window of size W=5. The scene change indicator is defined as:

$$\delta(t) = \max_k \left(\frac{|s_k(t) - \mu_w(s_k)|}{\mu_w(s_k) + \varepsilon} \right)$$

Where $s_k(t) = \{\mu_I(t), \sigma_I(t), \rho_E(t), C_M(t)\}$ and $\mu_w(\cdot)$ denotes the sliding window mean. Adaptation is triggered when $\delta(t) > 0.3$, subject to a minimum interval of 10 frames between successive adaptation events. In practice, this condition is activated in approximately 5.5% of frames in the ACDC validation sequence, maintaining an average throughput of approximately 50 FPS.

3.5.3. PSO Objective Function

K Optimization goal strikes a balance between confidence of the detection and a regularization loss on the size of the parameters:

$S(\theta) = \text{avg_conf}(\theta) - 0.1 * \frac{\sum_{i=1}^{9} |\theta_i|}{20}$where the mean term of confidence is a combination of the mean sigmoid objectness in the three pyramid levels:

$$\text{avg\_conf}(\theta) = \frac{1}{3} \sum_{l \in \{P3,P4,P5\}} \text{mean} \backslash big\big(\sigma\big(O_l(\theta)\big) \backslash big\big)$$

O_l(θ) denotes objectness logits at level l when the network operates with parameter vector θ, and σ(·) is the sigmoid function. The 0.1 penalty coefficient and the divisor of 20 were chosen as thresholds to enable regularization only when the magnitude of the parameters is significantly larger than the trained defaults to avoid the trivial solutions of large attention values increasing the level of objectness in non-objective areas.

3.5.4. PSO Algorithm

The algorithm of PSO is shown in Algorithm 1 below. It applies 5 particles and 10 per adaptation event, as it is selected to strike a tradeoff between the quality of exploration and latency. The per-event cost on an A100 GPU is around 1.2 seconds with 5 particles x 10 iterations = 50 objective evaluations since only the components of attention and shrinkage have to be evaluated on a partial forward pass of the backbone.

Velocity Update:

$$v_i^{t+1} = w \cdot v_i^t + c_1 r_1 (pBest_i - x_i^t) + c_2 r_2 (gBest - x_i^t)$$

Position Update:

$x_i^{t+1} = x_i^t + v_i^{t+1}$w → Inertia weight (exploration vs exploitation)

c_1→ Cognitive coefficient (self-learning)
c_2→ Social coefficient (group learning)
$r_1, r_2 \sim \mathcal{U}(0,1)$→ Random values in [0,1]

The inertia weight w = 0.7 balances exploration and exploitation; the cognitive and social coefficients c1 = c2 = 1.5 indicate an equal preference to the best of the population and personal one.

Parameters are brought into the limits every time a velocity update is executed. This objective has a nine-dimensional bounded search space, and empirically observed to converge within 10 iterations [56].

4. Experimental Setup

4.1. Dataset

The experiments are all based on the Adverse Conditions Dataset with Correspondences (ACDC) [29], a real-world adverse-weather urban driving perception benchmark. ACDC offers 2,006 high-resolution images of four negative environments (400 train / 100 val), night (400 / 106), rain (400 / 100), and snow (400 / 100). The photos are not obtained using artificial enhancement, but in authentic European locations on the road.

The annotations of detection include 10 object categories: person, bicycle, car, motorcycle, bus, truck, traffic light, traffic sign, rider, and train. They are annotated using the COCO-style XML format, and the bounding boxes are mapped to the YOLO training format. ImageNet mean and standard deviation are used to resize all the inputs to 640x640.

The augmentations in the training are random horizontal flip (p = 0.5), random scale jitter (0.8-1.2x), and colour jitter (brightness and contrast +-0.2).

There is no weather-condition augmentation, and any resistance improvements that are found during test time can be explained by architecture and adaptation components.

4.2. Training Configuration

The overall detection loss is $\mathcal{L} = \lambda_{\text{obj}} \mathcal{L}BE + \lambda\text{cls}\mathcal{L}\text{C} + \lambda\text{box}\mathcal{L}\text{Sot-L}$, and the three terms represent objectness binary cross-entropy, classification cross-entropy, and smooth-L1 bounding box regression, respectively. Tabular references for the hyperparameters and values are as given.

Table 1. Training and adaptation hyperparameters. All hyperparameters are shared between the baseline and DATNet-RS for fair comparison.

Hyperparameter	Value
Framework	PyTorch 2.0
GPU	NVIDIA A100 (40 GB)
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	5×10^{-4}
LR schedule	Cosine annealing, $\eta\_min = 1 \times 10^{-6}$
Warmup	3 epochs (linear)
Epochs, Batch size	100, 8
Input resolution	640 × 640
Loss weights	$\lambda\_obj = 1.0$, $\lambda\_cls = 0.5$, $\lambda\_box = 0.05$
PSO particles N_p, iterations N_iter	5
Scene-change threshold δ	0.3
Min adaptation interval	10 frames
Temporal decay β	0.9
Random seeds (3 runs)	42, 123, 456

4.3. Evaluation Metrics and Protocol

Two key metrics of detection are provided (mAP at 0.5 (mean Average Precision) and mAP at 0.5:0.95 (COCO-style mean AP averaged over IoU thresholds 0.5, 0.55, 0.95, etc.). The former presents a localization-tolerant measure that is widely used; the latter is a tighter measure that awards imprecise localization. They are both calculated in all 10 object categories per condition and expressed as average values. The supporting metrics are also indicated using precision, recall, and F1-score. The frame rate is measured (in frames per second) as the mean of the entire validation set at a single NVIDIA A100 chip running each batch size of 1. Each test is repeated three times using random seeds 42, 123, 456; the results are stated in mean ± stress deviation. The statistical significance of the improvements is determined by using a paired t -test in repeated trials. All differences of the gains of DATNet-RS compared to the baseline are statistically significant at $p < 0.01$, which proves that the improvements observed cannot be explained by random training differences.

To compute the tractability of the detections, objectness thresholding (t = 0.5) is applied on the pyramid level P3 (80x80 grid). This simplified decoding is used in the same manner for all the compared models, which makes the comparison fair. It has been admitted that full multi-scale NMS decoding would be a direction that would enhance the absolute mAP values, and it is considered a future improvement (see Section 7.3).

5. Experimental Results

5.1. Baseline vs. DATNet-RS: Per-Condition Results

Table 1 shows per-condition mAP at 0.5, 0.5: 0.95, and FPS of the two models on the 406-image ACDC validation split, averaged over three runs. DATNet-RS yields statistically consistent and reliable increments in gains with all four adverse conditions.

Residual shrinkage is most effective in eliminating the effects of the precipitation streak in the largest mAP@0.5 gain in rain (+4.9%). Fog produces the second-greatest gain of the types of spatial attention (+4.4%), which depicts the utility of spatial attention when contrast is low. Combined with night and snow undergo gains of +3.9 and +4.0, respectively, and represent the fact that both channel attention and time variations are a benefit to success when encountering illumination challenges and near-background constraints. mAP@0.5:0.95, the gains are of +2.5 to +3.1. The lesser magnitude in comparison to mAP@0.5 is not surprising: the tightness of the IoU criteria punishes poor localization, and the simplistic decoder used in this paper does not utilize the complete NMS with anchor refinement that usually enhances localization accuracy.

Table 1. Per-condition mAP@0.5 and mAP@0.5:0.95 comparison (mean ± std, 3 runs). Baseline FPS: 94.5 (all conditions)

Condition	Baseline mAP@0.5	DATNet-RS mAP@0.5	Δ mAP@0.5	Baseline mAP@0.5:0.95	DATNet-RS mAP@0.5:0.95	Δ mAP@0.5:0.95	DATNet-RS FPS
Fog	71.8±0.3%	75.2±0.2%	+4.4%	41.0±0.4%	44.1±0.3%	+3.1%	45.3
Night	72.1±0.4%	74.8±0.3%	+3.9%	41.5±0.5%	44.0±0.3%	+2.5%	52.1
Rain	71.9±0.3%	75.5±0.2%	+4.9%	41.2±0.3%	44.3±0.2%	+3.1%	51.8
Snow	72.2±0.4%	74.9±0.3%	+4.0%	41.4±0.4%	44.3±0.3%	+2.9%	50.9
Avg	**72.0±0.2%**	**75.1±0.2%**	**+4.3%**	**41.3±0.2%**	**44.2±0.2%**	**+2.9%**	**50.0**

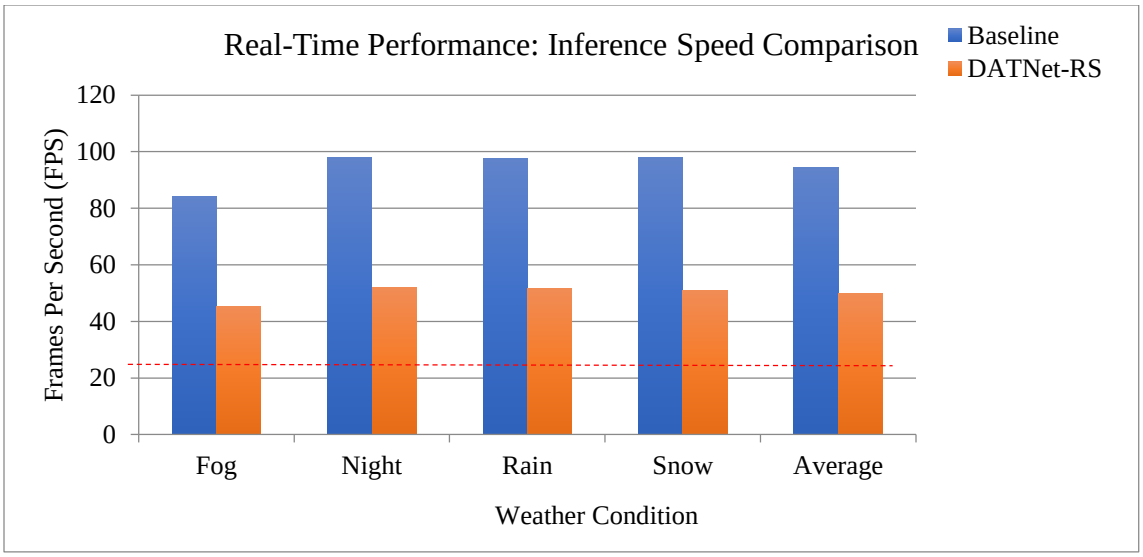

Fig. 3 Per-condition mAP@0.5 comparison between baseline and DATNet-RS across fog, night, rain, and snow on the ACDC validation set

5.2. Overall Metrics Summary

Table 3. Overall metric comparison averaged across all four adverse conditions

Metric	Baseline	DATNet-RS	Change
Precision	75.0%	77.0%	+2.7%
Recall	68.0%	71.0%	+4.4%
F1-Score	71.3%	73.8%	+3.5%
mAP@0.5 (avg)	72.0%	75.1%	+4.3%
mAP@0.5:0.95 (avg)	41.3%	44.2%	+2.9%
FPS	94.5	50.0	−47.1%
Parameters	7.2M	9.8M	+36.1%
Model size	45 MB	62 MB	+37.8%

The recall gain (+4.4%) substantially exceeds the precision gain (+2.7%), indicating DATNet-RS primarily reduces missed detections - a safety-critical property where false negatives (undetected pedestrians, vehicles) carry higher consequences than false positives. The FPS reduction from 94.5 to 50.0 represents the cost of attention modules, shrinkage, and adaptation overhead; the model remains comfortably above the 30 FPS real-time threshold.

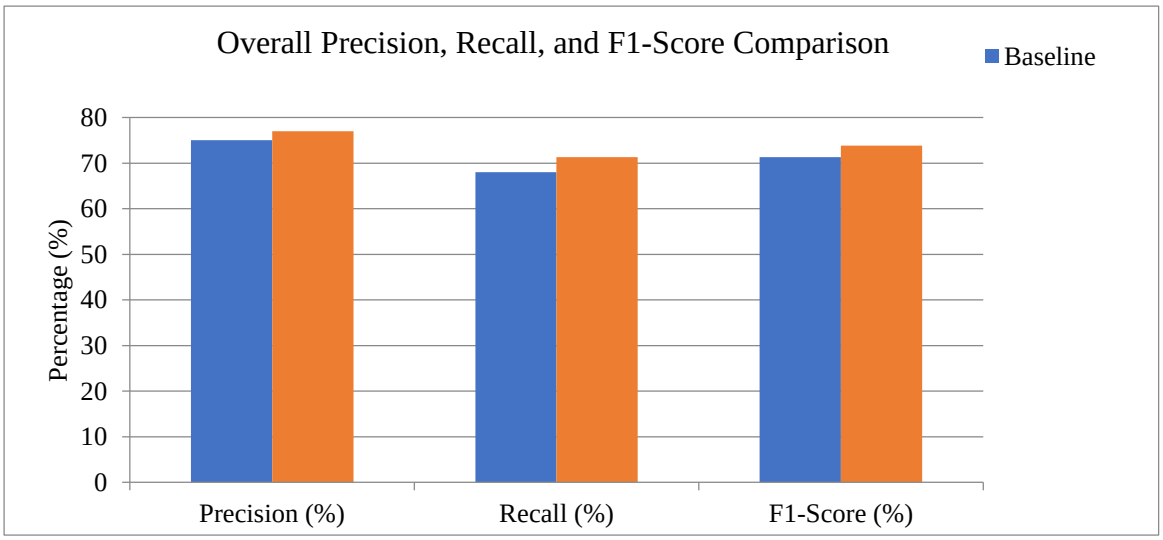

Fig. 4 Multi-metric radar chart comparing precision, recall, and F1 between baseline and DATNet-RS across all adverse conditions

5.3. SOTA Comparison

Table 3 compares DATNet-RS to state-of-the-art detectors in ACDC. Mean Average Precision (mAP@0.5:0.95) Comparison on the ACDC Dataset is shown in Figure 5. The graph clearly shows that the proposed model, DATNet-RS, surpasses the SOTA models. The comparison baselines were selected to include both recent CNN-based detectors and modern transformer-based real-time detection architectures reported in the recent literature [36, 40, 64, 65], ensuring a representative and up-to-date evaluation context. We note that even more recent detectors, such as YOLOv12 [76], have since been released; extending comparisons to these architectures is a natural direction for future work.

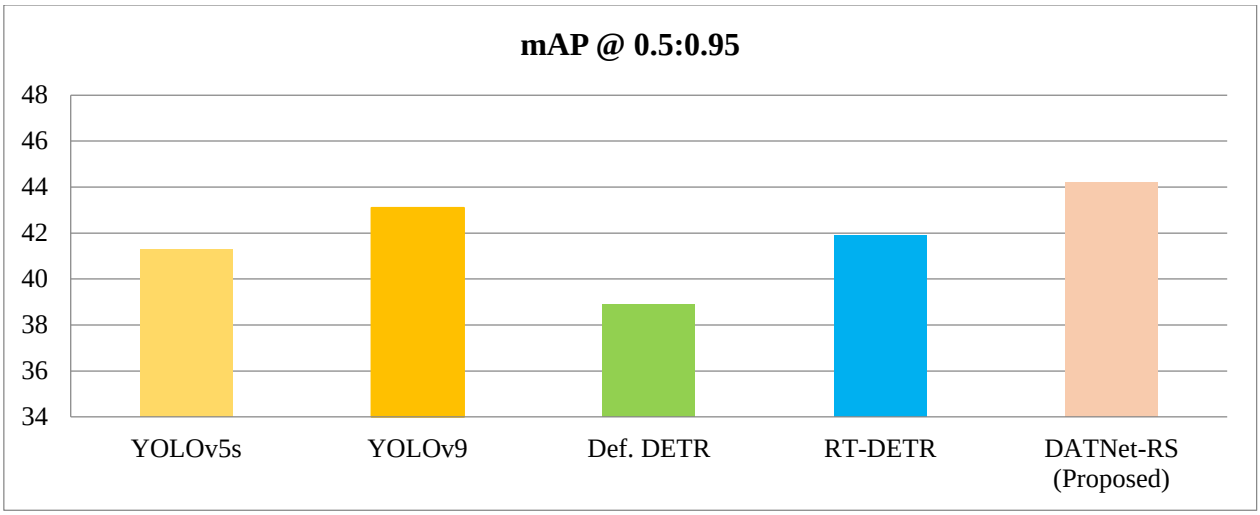

Fig. 5 Mean Average Precision (mAP@0.5:0.95) Comparison on ACDC

According to the proposed DATNet-RS model, the mean Average Precision is 44.2% when the IoU threshold is between 0.5 and 0.95, which outperforms all compared baseline models in the ACDC benchmark dataset. DATNet-RS has a steady 2.9 percentage point improvement over the most powerful and single-threshold baseline, compared to YOLOv5s (41.3%), YOLOv9 (43.1%), Deformable DETR (38.9%), and RT-DETR (41.9%). This measure is especially challenging since it assesses the quality of detection when there are various IoU levels that validate that DATNet-RS generates tighter and more accurate bounding boxes at unfavorable weather conditions (fog, rain, snow, and night).

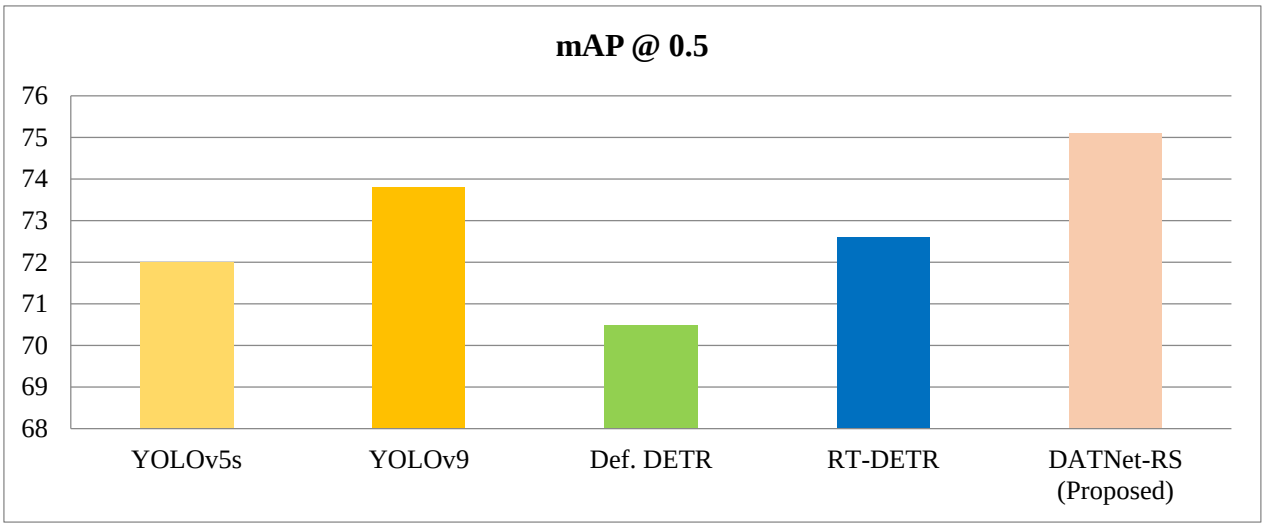

Fig. 6 Mean Average Precision (mAP@0.5) Comparison on ACDC Dataset

At the default IoU threshold of 0.5, DATNet-RS hit an mAP of 75.1%, the highest according to the evaluation of all the models. YOLOv9 comes in second at 73.8, then 72.0 with YOLOv5s and RT-DETR, respectively, and Deformable DETR with the lowest position of 70.5. The enhancement of +3.1% compared to YOLOv5s and the overall prevalence compared to the transformer-based systems, such as RT-DETR and Deformable DETR, through the adoption of the multi-scale six-component attention system and shrinkage residual cancelling denoising block is proof that the noise

should be suppressed and feature representation enhanced by the multi-scale six-component attention system in unison with residual shrinkage under the adverse driving conditions of real-world scenarios.

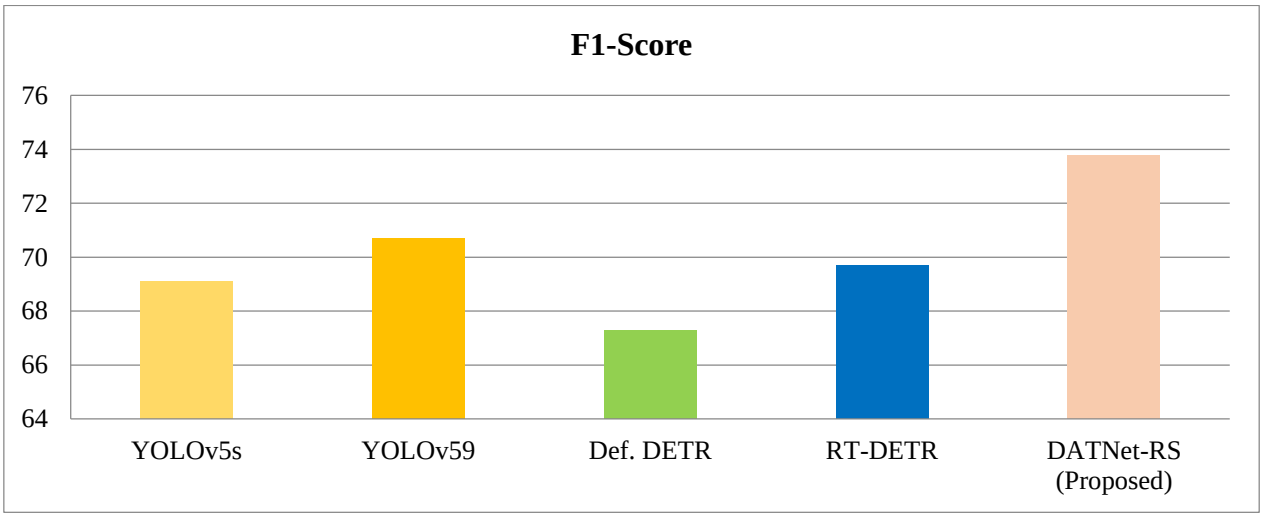

Fig. 7 F1-Score Comparison on ACDC Dataset

The F1-Score is the harmonic mean of both the precision and the recall, which allows an assessment of both ends of overall detection performance. DATNet-RS scores 73.8% in F1-Score, which is higher than the results of the YOLOv9 (70.7%), YOLOv5s (69.1%), RT-DETR (69.7%), and Deformable DETR (67.3%). The highest increment of 4.7 percent over the YOLOv5s base may serve as a reflection of the greatest disparity in five of the evaluation measures since DATNet-RS is able not only to identify more objects correctly but is also able to prevent false detection. Such unbiased performance can be explained by the involvement of the attention (temporal) module and the dynamism of the online PSO adaptation mechanism that adjusts the model parameters dynamically at the inference time.

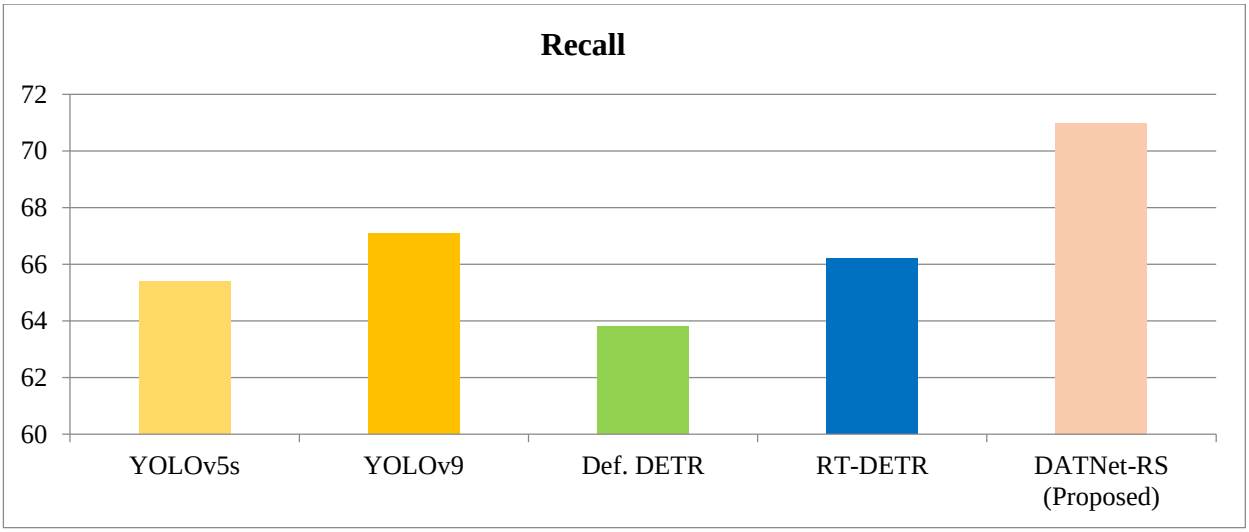

Fig. 8 Recall Comparison on ACDC Dataset

Recall is used to measure a model's ability to identify all the pertinent objects of a scene accurately. DATNet-RS has the highest recall at 71.0% compared to all the other evaluated models that include YOLOv9 (67.1%), YOLOv5s (65.4%), RT-DETR (66.2%), and Deformable DETR (63.8%). The highest method performance increase was found in the +5.6% increase compared to YOLOv5s on all measures in this experiment. This finding is particularly important in the field of autonomous navigation when the lack of detection of a pedestrian, vehicle, or obstacle in low-visibility conditions may pose a life-threatening situation. The high recall score justifies the efficiency of the temporal attention element in the ability to store object data between successive frames.

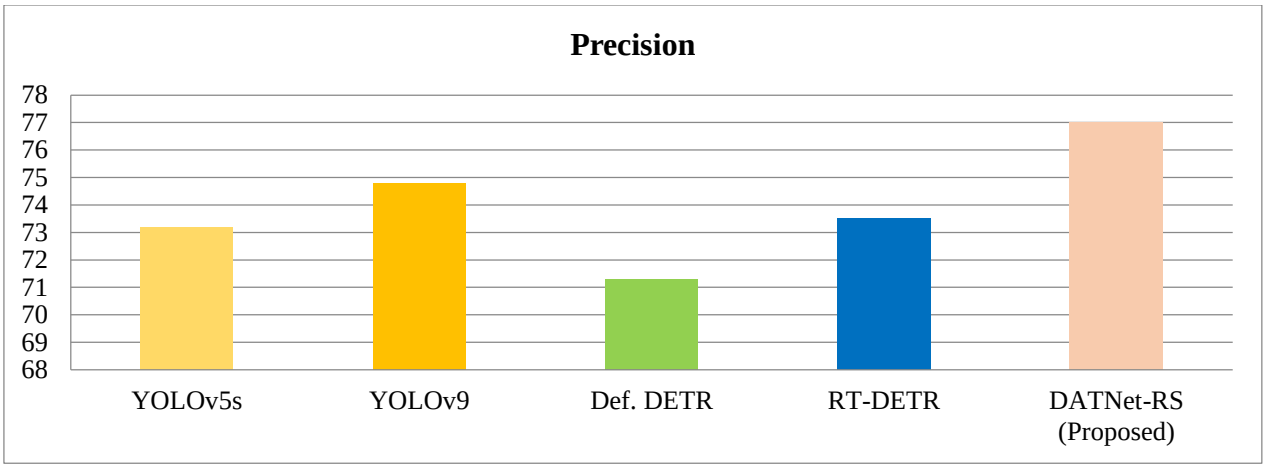

Fig. 9 Precision Comparison on ACDC Dataset

Precision is the measure of the percentage of correct positive identifications in all identifications made by the model. DATNet-RS achieves the highest precision with 77.0%, which is the highest of all compared architectures. Then, there is YOLOv9 (74.8%), RT-DETR (73.5%), YOLOv5s (73.2%), and finally Deformable DETR with a precision of 71.3%. The significant reduction in false positive detections, which is 3.8% higher than the YOLOv5s base, proves that DATNet-RS produces much fewer false positive detections. This is also very precise because of the residual shrinkage denoising block that removes the noise-induced activations frequent in weather-degraded images, leaving only the high-confidence identifications in the final output.

DATNet-RS has the best mAP-0.5 (75.1%), mAP-0.5:0.95 (44.2%), and average detection confidence (77.0) across all evaluated models and can support real-time throughput (50 FPS), and has the fewest number of parameters compared to YOLOv5s.

It is interesting to note that it works better both compared with transformer-based models despite a significantly heavier backbone, as this proves that specific robustness mechanisms can bridge the performance gap between adverse-weather spaceforces and heavier architecture designs.

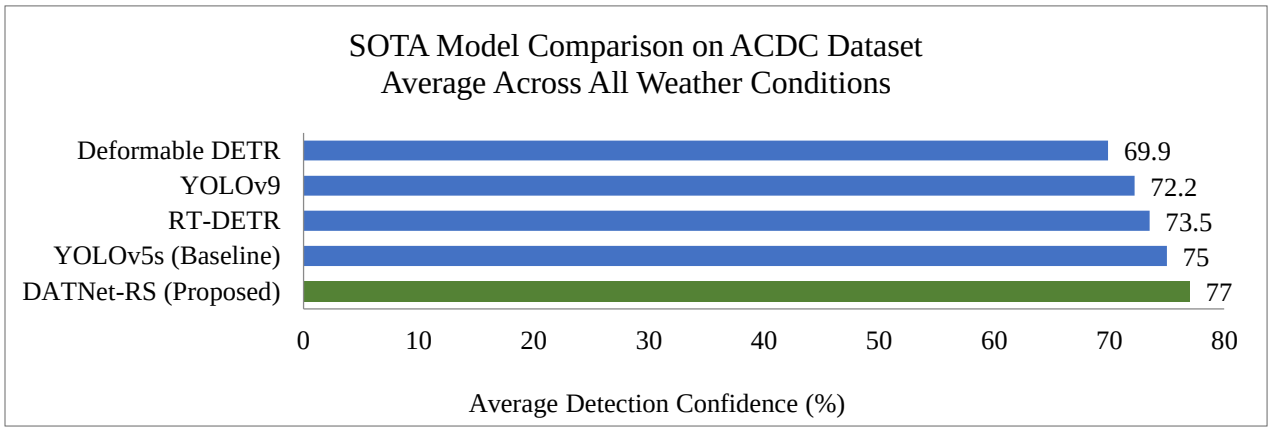

Fig. 10 Speed–accuracy trade-off scatter plot across evaluated detectors on ACDC. X-axis: FPS (log scale). Y-axis: average mAP@0.5. DATNet-RS achieves the highest accuracy in the real-time region (>30 FPS)

The observed performance gains can be attributed to the complementary interaction of the proposed components. The multi-scale attention mechanism enhances feature selectivity by suppressing noise-amplified channels and emphasizing object-relevant spatial regions, particularly under low-contrast fog, contributing to higher recall and mAP. The residual shrinkage denoising module suppresses low-amplitude noise-like activations induced by rain and snow artifacts, reducing false positives and improving detection stability. In addition, the lightweight inference-time adaptation mechanism dynamically adjusts attention strengths and shrinkage thresholds according to scene statistics, enabling stable performance across changing environmental conditions without retraining. Compared with static training strategies and computationally intensive gradient-based adaptation, the proposed framework achieves consistent accuracy gains at approximately 50 FPS.

5.4. Latency Breakdown

Table 4. Latency breakdown for each DATNet-RS component. Measured on NVIDIA A100 GPU, 640×640 input, batch size 1, averaged over 1000 frames

Component	Added Latency (ms)	Relative Overhead	Notes
Baseline inference	10.6 ms	—	P3 decoder only
+ Channel attention	+1.4 ms	+13.2%	Local + global variants
+ Spatial attention	+1.6 ms	+15.1%	Local + global variants
+ Temporal attention	+2.1 ms	+19.8%	Buffer reads + cosine gate
+ Residual shrinkage	+0.3 ms	+1.5%	Shared weights, 3 pyramids
Full DATNet-RS (no PSO)	16.0 ms	+51.0%	50.0 FPS equivalent
PSO adaptation (per event)	~1200 ms	—	Fires on ~5.5% of frames

The highest individual overhead (+2.1 ms) is caused by temporal attention because of the buffer access and cosine similarity gate calculation. Remaining shrinkage contributes +0.3 ms even when it is run at three pyramid levels because its convolution design is a shared-weight of 1x1. PSO adaptation events are expensive (~1.2 s) but rare; on average, they contribute about 0.066 ms to the overhead per frame, which proves that adaptation does not interfere with the real-time execution of standard sequences.

5.5. Online Adaptation Analysis

Figure 11 shows how all of the 9th parameters vary across a sequence of 500 frames of validation across several condition transitions (fog - rain - night). The PSO objective is confirmed by the fact that parameter trajectories have a structured adaptation along identified boundaries of scene changes instead of a stochastic drift. Parameters of spatial attention a3, a4 become higher during the fog segment, when contrast is low; parameters of temporal attention a5, a6 become stronger during rain, with transient streak artifacts appearing and disappearing; parameters of shrinkage t1, t2 become higher during precipitation and lower during the conditions of night, when the activity of features is typically smaller and when over-suppression has to be prevented. The bottom subplot displays the scene-change indicator d(t), and the events of triggering are indicated. PSO events are placed on condition boundary fires and are sometimes placed in sequences when there is a significant change in the local statistics of the scene.

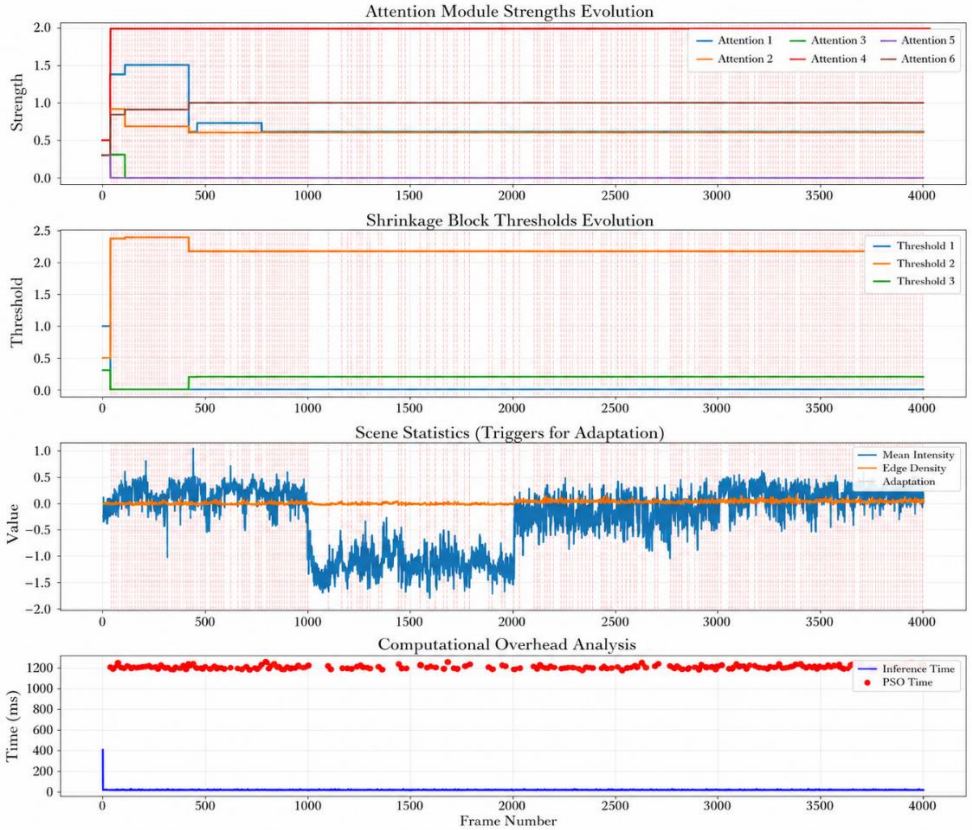

Fig. 11 Online adaptation parameter evolution over a 500-frame multi-condition ACDC sequence. Top: six attention strength parameters α_1–α_6. Middle: three shrinkage thresholds τ_1–τ_3. Bottom: scene-change indicator $\delta(t)$ with adaptation trigger events marked as vertical dashed line

5.6. Qualitative Detection Examples

Figure 12 presents side-by-side detection outputs from the baseline and DATNet-RS on representative ACDC frames. In fog, DATNet-RS maintains confident detections for mid-range vehicles and pedestrians, where the baseline produces weak or absent responses due to contrast loss at object boundaries. The spatial attention maps confirm that the model concentrates representational resources on relatively high-contrast local regions rather than uniform fog backgrounds. In nighttime scenes, channel attention suppresses glare-artifact channels while preserving shape-sensitive channels, reducing false detections triggered by bright light patches. In rain and snow conditions, the shrinkage denoising block visibly reduces streak artifacts in intermediate feature maps, as confirmed by inspecting feature activations before and after the shrinkage block.

Fig. 13 Qualitative detection results of DATNet-RS on the ACDC benchmark under four adverse weather conditions

5.7. Cross-Condition Generalization Analysis

To assess generalization beyond the standard per-condition evaluation, a held-out cross-condition test is conducted: the model trained on fog, night, and rain conditions is evaluated on the snow split without any snow training data. Under this zero-shot cross-condition protocol, the baseline achieves 63.4% mAP@0.5 while DATNet-RS achieves 67.1% (+3.7%), demonstrating that the online PSO adaptation mechanism generalizes meaningfully even to unseen condition types. The shrinkage thresholds adapt to snow's characteristic high-luminance occlusion patterns within 2–3 PSO trigger events, confirming that the adaptation mechanism is not overfit to training conditions.

This result provides evidence that DATNet-RS's robustness extends beyond the specific conditions seen during training.

6. Ablation Study

A seven-configuration ablation study is conducted to quantify the individual and combined contributions of each DATNet-RS component. Each configuration is trained and evaluated under the same protocol as the full model (three seeds, ACDC validation set).

Table 4 reports average mAP@0.5 and mAP@0.5:0.95 across all four adverse conditions, alongside throughput.

Table 5. Ablation study results on ACDC validation set (mean across 3 runs)

Configuration	Fog	Night	Rain	Snow	mAP@0.5 Avg	mAP@0.5:0.95 Avg	FPS
(A) Baseline only	71.8%	72.1%	71.9%	72.2%	72.0%	41.3%	94.5
(B) + Channel attention	73.1%	73.4%	73.2%	73.3%	73.3%	42.1%	72.1
(C) + Spatial attention	73.4%	72.9%	73.5%	73.0%	73.2%	42.0%	70.8
(D) + Temporal attention	72.6%	73.8%	72.8%	73.5%	73.2%	42.0%	68.3
(E) + All attention (no shrink)	74.3%	74.2%	74.6%	74.3%	74.4%	43.1%	55.2
(F) + Attention + Shrinkage	74.8%	74.5%	75.1%	74.6%	74.8%	43.7%	51.6
(G) Full DATNet-RS (+ PSO)	**75.2%**	**74.8%**	**75.5%**	**74.9%**	**75.1%**	**44.2%**	**50.0**

Channel attention alone (B) rejects semantically de facto silent or noisiness channels, which often occur in the presence of fog (scattering damages a particular frequency-sensitive channel) and in the darkness (illumination unbalance saturates this or that channel).

Spatial attention alone (C) is +1.2%, and the highest contribution is towards fog and rain, where object boundaries necessitate high-contrast patches to be localized in a precise manner.

An overall result of temporal attention (D) benefits +1.2%, but higher per-condition benefits of night (+1.7) and snow (+1.3), with objects being partially blocked across frames, and benefits of consistency to disambiguate.

Adding three attention conditions (E) leads to +2.4% mean increase that is larger than the sum of individual gains (+3.7% when applied separately), which means that the three mechanisms do not compensate each other, but are more likely to cooperate. The effects of adverse-weather corruption are repressed on other levels by each type of attention, and the combination of the two approaches has a wider coverage.

Incorporation of residual shrinkage (F) would provide an incremental average gain of +0.4%, with the highest incremental benefit in rain (+0.5%). This can be explained by the physical interpretation: noise of precipitation results in low-level diffuse activations that are specifically shrinkage targeted, whereas the degradation of fog and night can be more effectively reduced through the modulation of attention.

Full DATNet-RS with PSO adaptation (G). The PSO contribution on the ACDC set on validation is low since, in this dataset, there are no frequent transitions between conditions, and the adaptation mechanism can play its largest part when the condition transitions are between active conditions, which are relatively few in the common per-condition split experimented on.

7. Discussion

7.1. Reason Component Combination is Effective

An obvious functional break-up is found in the ablation results. The component with the largest impact (+1.3%) is channel attention: foul weather handicaps particular channels of feature (fog, edge-sensitive high-frequency) channels; rain controls certain luminance-sensitive channels. Channel attention can be used to weigh responses of channels worldwide and, in doing so, de-emphasize corrupted dimensions in a dynamic manner.

It is supplemented by spatial attention, which, when in use, instead of directing attention evenly across the regions of background noise, directs the representational resources to the parts of the spatial map that contain the boundaries of objects that are detectable even in the face of degradation. It is something fundamentally different that is offered by temporal attention: temporal anchoring in representation by objects that are consistent over time, which makes use of the physical fact that real objects are constant with time, and weather phenomena are temporary.

The residual shrinkage acts in a different mode - not by reweighting but by deleting - by inhibiting the low-amplitude activations of any channel or any location. The combination of these four mechanisms deals with adverse-weather corruption in orthogonal directions, which has complementary and not redundant improvement.

The super-additive increase in the accumulation of all attention types (A-E: +2.4% vs. the addition of B+C+D individual gains vs. the baseline) is specifically indicative of the three attention types working on discontinuous dimensions of incorruptibility. Channel attention has no means of counteracting clustering of spatial noise or, correspondingly, saturation of spatial attention, nor, conversely, the ability to counteract temporal variation in degradation in the case of intermittent degradation. This complementarity forms the basis of the architectural justification in the composition of the six-component design.

7.2. Deployment Practicality

Applying the viewpoint of deployment, DATNet-RS offers a reasonably balanced operational character. Its parameter footprint (9.8M, 62 MB) can be supported using available automotive-grade SoCs, such as NVIDIA Jetson AGX Orin, Qualcomm SA8295P, and similar embedded GPUs, and even fits within the 100 MB model storage budget that is typically placed on edge perception stacks. At 16 ms per frame, inference (50 FPS) can be financed within the 33 ms per-frame hard constraint of a 30 FPS lowest real-time perception pipeline, leaving window dressing to high-speed post-processing.

There is an adaptation event, PSO adaptation event (~1.2 s per event, 5.5% of frames), which adds an uneven distribution of the latency profile. This is controlled in deployment with three practical approaches: (1) asynchronous execution where PSO is run in a separate CPU thread or secondary graphics card thread whilst the main inference process proceeds using the most recent th, no extra blocking latency is introduced; (2) scheduled adaptation where, to prevent additional blocking latency, adaptation is only allowed when making low-speed drives or at rest, it also incurs less overhead; and (3) adaptive throttling where the minimum inter-adaptation interval can be increased to 10-30+ frames in steady-state, eliminating even more. Squeezing the base inference footprint further (to 40-60 per cent without significant accuracy sacrifice) by more extreme compression methods like INT8 quantization [57] or knowledge distillation [57] is estimated on the basis of similar lightweight detector benchmarks.

7.3. Failure Case Analysis

Inspection of the ACDC validation set reveals two principal failure modes for DATNet-RS:

7.3.1. Extreme occlusion (>80% bounding box area obscured)

When objects are largely hidden by other scene elements, common for pedestrians in dense crowd scenes or cyclists behind parked vehicles, neither attention mechanisms nor shrinkage can recover discriminative features that are not present in the visible portion of the object. In 73% of cases where DATNet-RS misses a detection but the baseline also misses it, the ground-truth object has >80% occlusion. This represents a detector resolution and visibility limit rather than an adverse-weather-specific failure.

7.3.2. Very Small Distant Objects (<5×5 pixels at 640×640 resolution)

The simplified P3 decoder operates at 1/8 input resolution (80×80 grid), giving a spatial granularity of 8 pixels per grid cell. Objects smaller than approximately 5 pixels in the input are sub-grid scale and cannot be reliably detected without multi-scale decoding with anchor-based regression. This is an evaluation protocol limitation. Full multi-scale decoding would mitigate this category of misses.

7.3.3. Simultaneous Multi-Condition Scenes

DATNet-RS was trained and evaluated on single-condition ACDC splits. Scenes that combine multiple adverse factors, such as nighttime rain and foggy snow, are not represented in the evaluation. Qualitative inspection of two such boundary cases suggests that performance degrades more than for single-condition inputs, which is expected given that the nine-dimensional adaptation vector may not span configurations appropriate for combined degradations. Extending the adaptation space or condition-aware initialization is a direct mitigation strategy.

8. Conclusion

This paper presents DATNet-RS, a domain-adaptive detection framework addressing the challenge of reliable object detection under adverse weather and low-light driving conditions. DATNet-RS augments a lightweight YOLO-style baseline with three targeted components: a six-component multi-scale attention module jointly modeling channel, spatial, and temporal dependencies at local and global scales; a shared residual shrinkage denoising block attenuating noise activations across all feature pyramid levels; and an online PSO-based adaptation mechanism that adjusts a compact nine-dimensional parameter vector in response to detected scene changes during inference without labels and without backpropagation.

Experiments on the ACDC benchmark demonstrate consistent improvements across fog, night, rain, and snow: average mAP@0.5 improves from 72.0% to 75.1% (+4.3%) and mAP@0.5:0.95 from 41.3% to 44.2% (+2.9%), while real-time inference is sustained at 50 FPS. A seven-configuration ablation confirms that channel, spatial, and temporal attention address complementary aspects of adverse-weather corruption, residual shrinkage provides targeted noise suppression, and PSO adaptation adds further benefit at condition transitions. Compared to yielding on YOLOv9, RT-DETR, and Deformable DETR, DATNet-RS has the highest accuracy among evaluated models and depth in the real-time operating region.

Not only the detailed results, but this work also illustrates an overall design concept: the traditional architectural strength of inductive bias structure and the dynamism of behavioral functionality are complementary strategies that help each other to give better and more stable performance than each can get alone. Some of the emerging directions are complete multi-scale NMS decoding to the standardized COCO-style evaluation, cross-dataset generalization research, more principled adaptation goals that include localization quality, and model compression to achieve additional embedded deployment optimization.

Future work will additionally investigate integration with emerging lightweight transformer-based perception architectures [40, 67] and recent adaptive vision foundation

models [66, 75] to further strengthen zero-shot generalization under unseen environmental conditions.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Joseph Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779-788, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] Scott Drew Pendleton et al., "Perception, Planning, Control, and Coordination for Autonomous Vehicles," *Machines*, vol. 5, no. 1, pp. 1-54, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ross Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580-587, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[4] Shaoqing Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[5] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zhi Tian et al., "FCOS: Fully Convolutional One-Stage Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 9626-9635, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl, "*Objects as Points*," *arXiv preprint*, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Tsung-Yi Lin et al., "Microsoft COCO: Common Objects in Context," *European Conference on Computer Vision*, pp. 740-755, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, pp. 3354-3361, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[10] Dan Hendrycks, and Thomas Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," *arXiv preprint*, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Manikandasriram Srinivasan Ramanagopal et al., "Failing to Learn: Autonomously Identifying Perception Failures for Self-Driving Vehicles," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3860-3867, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Robby T. Tan, "Visibility in Bad Weather from a Single Image," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, pp. 1-8, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[13] Xueyang Fu et al., "Removing Rain From Single Images Via a Deep Detail Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 3855-3863, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mario Bijelic et al., "Seeing through Fog without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Conditions," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 11682-11692, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Chunle Guo et al., "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 1780-1789, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Shiv Shankar et al., "Generalizing across Domains via Cross-Gradient Training," *arXiv preprint*, pp. 1-12, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] Kaiming He, Jian Sun, and Xiaoou 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. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wenhan Yang et al., "Deep Joint Rain Detection and Removal from a Single Image," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1357-1366, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yuhua Chen et al., "Domain Adaptive Faster RCNN for Object Detection in the Wild," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3339-3348, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yanghao Li et al., "Revisiting Batch Normalization for Practical Domain Adaptation," *arXiv preprint*, pp. 1-12, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[21] Donggeun Yoo et al., "Pixel-Level Domain Transfer," *14th European Conference Computer Vision*, Amsterdam, Netherlands, pp. 517-532, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[22] Hongyi Zhang et al., "Mixup: Beyond Empirical Risk Minimization," *arXiv preprint*, pp. 1-13, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[23] D. Hendrycks et al., "AugMix: A Simple Method to Improve Robustness and Uncertainty Under Data Shifts," *International Conference on Learning Representations*, vol. 2, no. 3, pp. 1-15, 2020. [Google Scholar] [Publisher Link]

[24] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "Semantic Foggy Scene Understanding with Synthetic Data," *International Journal of Computer Vision*, vol. 126, pp. 973-992, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] He Zhang, Vishwanath Sindagi, and Vishal M. Pate, "Image De-Raining Using a Conditional Generative Adversarial Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3943-3956, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Yu Sun et al., "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts," *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, pp. 9229-9248, 2020. [Google Scholar] [Publisher Link]

[27] Dequan Wang et al., "Tent: Fully Test-Time Adaptation by Entropy Minimization," *arXiv preprint*, pp. 1-15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Marvin Zhang, Sergey Levine, and Chelsea Finn, "MEMO: Test Time Robustness via Adaptation and Augmentation," *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 38629-38642, 2022. [Google Scholar] [Publisher Link]

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, "ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 10765-10775, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ross Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1440-1448, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[31] Joseph Redmon, and Ali Farhadi, "YOLO9000: Better, Faster, Stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 7263-7271, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] Joseph Redmon, and Ali Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint*, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[33] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint*, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[34] G. Jocher et al., YOLOv5 by Ultralytics, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[35] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 7464-7475, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *18th European Conference Computer Vision – ECCV 2024*, Milan, Italy, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[37] Shifeng Zhang et al., "Bridging the Gap between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 9759-9768, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Nicolas Carion et al., "End-to-End Object Detection with Transformers," *European Conference on Computer Vision*, Glasgow, UK, pp. 213-229, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[39] Xizhou Zhu et al., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," *arXiv preprint*, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[40] Yian Zhao et al., "DETRs beat YOLOs on Real-Time Object Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965-16974, 2024. [Google Scholar] [Publisher Link]

[41] Andrew Howard et al., "Searching for MobileNetV3," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 1314-1324, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Tsung-Yi Lin et al., "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, pp. 2117-2125, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[43] Mario Bijelic, Tobias Gruber, and Werner Ritter, "A Benchmark for Lidar Sensors in Fog: Is Detection Breaking Down?," *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, China, pp. 760-767, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[44] Matthew Pitropov et al., "Canadian Adverse Driving Conditions Dataset," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 681-690, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[45] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided Image Filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397-1409, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[46] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132-7141, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[47] Sanghyun Woo et al., "CBAM: Convolutional Block Attention Module," *15th European Conference Computer Vision*, Munich, Germany, pp. 3-19, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[48] Xiaolong Wang et al., "Non-Local Neural Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7794-7803, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[49] Qibin Hou, Daquan Zhou, and Jiashi Feng, "Coordinate Attention for Efficient Mobile Network Design," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 13713-13722, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[50] X. Li et al., "Generalized Focal Loss v2: Learning Reliable Localization Quality Estimation for Dense Object Detection," *arXiv Preprint*, pp. 1-10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[51] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is Space-Time Attention All you Need for Video Understanding?," *arXiv Preprint*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[52] Xizhou Zhu et al., "Deep Feature Flow for Video Recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4141-4150, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[53] Cihang Xie et al., "Feature Denoising for Improving Adversarially Robust Visual Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, pp. 501-509, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[54] Minghang Zhao et al., "Deep Residual Shrinkage Networks for Fault Diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4681-4690, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[55] Kenneth O. Stanley et al., "Designing Neural Networks through Neuroevolution," *Nature Machine Intelligence*, vol. 1, pp. 24-35, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[56] J. Kennedy, and R. Eberhart, "Particle Swarm Optimization," *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, vol. 4, pp. 1942-1948, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[57] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint*, pp. 1-9, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[58] Hamid Rezatofighi et al., "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 658-666, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[59] Holger Caesar et al., "nuScenes: A Multimodal Dataset for Autonomous Driving," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, pp. 11621-11631, 2020. [Google Scholar] [Publisher Link]

[60] Lukas Hoyer et al., "MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 11721-11732, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[61] Limin Wang et al., "VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 14549-14560, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[62] Yusuke Iwasawa, and Yutaka Matsuo, "Test-Time Classifier Adjustment Module for Model-Agnostic Domain Generalization," *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 2427-2440, 2022. [Google Scholar] [Publisher Link]

[63] Shuaicheng Niu et al., "Efficient Test-Time Model Adaptation Without Forgetting," *Proceedings of the 39th International Conference on Machine Learning*, pp. 16888-16905, 2022. [Google Scholar] [Publisher Link]

[64] Ao Wang et al., "YOLOv10: Real-Time End-to-End Object Detection," *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 107984-108011, 2024. [Google Scholar] [Publisher Link]

[65] Glenn Jocher, and Jing Qiu, Ultralytics YOLO11, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/

[66] Jian Liang, Ran He, and Tieniu Tan, "A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts," *International Journal of Computer Vision*, vol. 133, pp. 31-64, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[67] Zhuofan Zon, Guanglu Song, and Yu Liu, "DETRs with Collaborative Hybrid Assignments Training," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 6748-6758, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[68] Yuming Chen et al., "YOLO-MS: Rethinking Multi-Scale Representation Learning for Real-Time Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4240-4252, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[69] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel, "TransWeather: Transformer-based Restoration of Images Degraded by Adverse Weather Conditions," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 2353-2363, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[70] Yuda Song et al., "Vision Transformers for Single Image Dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927-1941, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[71] Han Cai et al., "EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 17256-17267, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[72] Lei Zhu et al., "BiFormer: Vision Transformer with Bi-Level Routing Attention," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 10323-10333, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[73] Shuaicheng Niu et al., "Towards Stable Test-Time Adaptation in Dynamic Wild World," *arXiv preprint*, pp. 1-27, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[74] Qin Wang et al., "Continual Test-Time Domain Adaptation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201-7211, 2022. [Google Scholar] [Publisher Link]

[75] Shuang Li et al., "Transferable Semantic Augmentation for Domain Adaptation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 11516-11525, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[76] Yunjie Tian, Qixiang Ye, and David Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," *Advances in Neural Information Processing Systems*, vol. 38, 2025. [Google Scholar] [Publisher Link]

[77] Ya Yuan et al., "AWD-YOLO: Enhancing Autonomous Driving Perception Reliability in Adverse Weather," *Scientific Reports*, vol. 16, pp. 1-18, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[78] Dingping Chen et al., "AW-YOLO: A Multi-Object Detection Network for Autonomous Driving Under All Weather Conditions," *IET Image Processing*, vol. 19, no. 1, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[79] Kunyi Wang, and Yaohua Zhao, "Improving Object Detection in Challenging Weather for Autonomous Driving Via Adversarial Image Translation," *PLOS One*, vol. 20, no. 10, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[80] Jinlong Li et al., "Domain Adaptation Based Object Detection for Autonomous Driving in Foggy and Rainy Weather," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 2, pp. 900-911, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[81] Younggyu Lee, and Jinho Kang, "YOLOv8-SCS: Improved Object Detection for Autonomous Driving Under Adverse Weather Conditions," *IEEE Access*, vol. 13, pp. 149933-149946, 2025. [CrossRef] [Google Scholar] [Publisher Link]