Journal of Technology and Informatics (JoTI)
Vol. 8, No. 2, Bulan Oktober Tahun 2026
P-ISSN 2721-4842
E-ISSN 2686-6102

Condition-Aware Cross-Scale Transformers for Object Detection under Compound Adverse Environments

Fiston Oshasha Oshasha[1,2*], Saint Jean Djungu[1,2,3], Franklin Mwamba Kande[2,4]
[1]General Commissariat for Atomic Energy, Regional Center for Nuclear Studies of Kinshasa, P.O. Box 868, Kinshasa, DRC
[2]Center for Research in Applied Computing, University of Kinshasa, DRC
[3]Faculty of Science and Technology, University of Kinshasa, DRC
[4]Health Sciences Research Institute, Kinshasa, Democratic Republic of the Congo
e-mail: fiston.oshasha.oshasha@cgea-rdc.org[1*], saintjean.djungu@cria-unikin.net[2], franklin.mwamba@irss.cd[3]

Article Information

Article History:
Received : 6 June 2026
Revised : 21 October 2026
Accepted : 27 October 2026
Published : 30 October 2026

***Correspondence:**
fiston.oshasha.oshasha@cgea-rdc.org

Keywords:
adverse weather detection, autonomous driving, condition-aware adaptation, curriculum learning, object detection, transformer architecture

Copyright © 2026 by Author.
Published by Universitas Dinamika.

This is an open access article under the CC BY-SA license.

10.37802/joti.v8i2.1440

Journal of Technology and Informatics (JoTI)
P-ISSN 2721-4842
E-ISSN 2686-6102
https://e-journals.dinamika.ac.id/index.php/joti

***Abstract*:**

Object detection under adverse environmental conditions remains a critical challenge for autonomous driving systems, with performance degrading by over 40% under dense fog, heavy precipitation, or nighttime scenarios. Existing approaches either specialize for single degradation types or remain fundamentally condition-agnostic, failing to address the compound nature of real-world adverse conditions. We present a Condition-Aware Cross-Scale Transformer (CACST), a unified framework integrating explicit environmental condition estimation, adaptive image preprocessing, and cross-scale feature interaction. Our approach introduces: (1) a condition estimation module quantifying degradation parameters via supervised and contrastive learning; (2) a gated image adaptation mechanism with learnable ISP operations dynamically modulated by estimated conditions; and (3) an Enhanced Density-aware Cross-Scale Transformer (EDST) performing global attention across feature pyramid scales with spatial reliability weighting. Comprehensive experiments across five benchmarks demonstrate state-of-the-art performance: CACST achieves +4.1% mAP on ACDC (47.8% vs. 43.7%), +3.2% on Foggy Cityscapes (49.7% vs. 46.5%), and +6.7% on ExDark (74.9% vs. 68.2%) compared to the strongest specialized baseline, maintaining balanced accuracy across all conditions.

INTRODUCTION

Robust object detection under adverse environmental conditions represents a critical challenge for autonomous systems. While deep learning-based detectors achieve remarkable performance under controlled conditions [1]–[3], their accuracy degrades dramatically under

atmospheric obscurants and illumination variations. Empirical studies consistently demonstrate mAP reductions exceeding 30–40% under dense fog, heavy precipitation, or nighttime conditions [4], [5], severely limiting reliability in safety-critical domains such as autonomous driving and advanced driver assistance systems (ADAS). The fundamental difficulty arises from the compound nature of real-world adverse conditions. Deployed systems must handle fog scattering, rain distortion, low illumination, and sensor noise at once. Existing approaches fall into three limited categories: (1) multimodal sensor fusion [5], [6], which needs costly hardware; (2) condition-agnostic transformers [7], [8]; and (3) image-level preprocessing [9], [10] built for a single degradation. Previous research has already applied YOLO, SSD, and CNN detectors to access control, visitor counting, expression recognition, and gesture interfaces [11], [12], [13], [14]. Those systems assume a clearly visible target under ordinary lighting and do not model compound fog, rain, and night.

This motivates our central question: Can we design a unified detection framework that explicitly estimates environmental conditions, adaptively preprocesses degraded imagery, and performs degradation-robust multi-scale feature extraction, all within a single end-to-end architecture? We present CACST, addressing three core challenges: explicit degradation quantification, dynamic preprocessing modulation, and enhanced multi-scale feature interaction for small distant objects.

Contributions

1. Unified condition-aware detector. CACST estimates the environment, adapts the image with gated ISP operations, and fuses pyramid scales in one end-to-end model.
2. EDST. Cross-scale self-attention with density-aware spatial weights recovers small objects under scattering.
3. Gated ISP. Dehazing, denoising, tone mapping, and color correction are mixed by gates that depend on the estimated condition.
4. Curriculum on synthetic compounds. Training moves from light degradations to full compound weather.
5. Validation. CACST gains +4.1% mAP on ACDC, +3.2% on Foggy Cityscapes, and +6.7% on ExDark over the strongest specialized baseline.

METHOD

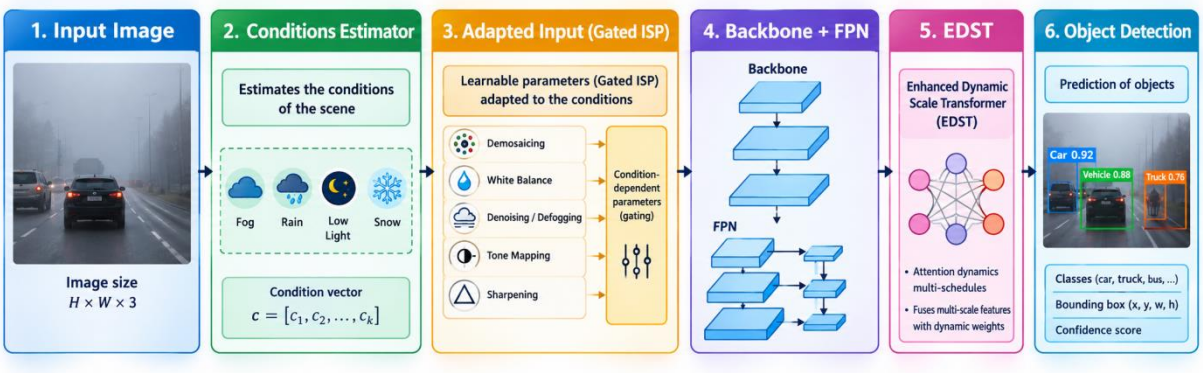

Figure 1. Overview of the proposed CACST framework. The condition estimation module first infers a latent condition vector z_x, which drives the gated ISP adaptation and modulates the density-aware cross-scale transformer for joint end-to-end optimization.

A. Condition Estimation Module

Explicit modeling of environmental conditions is crucial for adaptive processing. Our condition estimation module consists of a lightweight ResNet-18 encoder that processes the input image through global average pooling to produce a condition embedding:

$$z_c = \phi_c(I) = \text{GlobalPool}\big(\text{ResNet18}(I)\big) \in \mathbb{R}^{d_c} \quad (1)$$

where $d_c = 256$ in our implementation. The embedding z_c captures both discrete condition categories (clear, fog, rain, night) and continuous severity levels.

Supervision Strategy. During training, we supervise z_x using two complementary losses.

(1) Classification Loss. For labeled data with known conditions c_{gt}, we apply cross-entropy:

$$L_{cls} = -\sum_{k=1}^{K}\big(c_{gt} = k\big)\log p_k(z_c) \quad (2)$$

(2) Contrastive Loss. To learn continuous severity representations, we adopt a contrastive formulation inspired by SimCLR:

$$L_{contrast} = -\log\frac{\exp(\text{sim}(z_c, z_c^+)/\tau)}{\sum_j \exp(\text{sim}(z_c, z_j)/\tau)} \quad (3)$$

where z_x^+ represents embeddings of similar conditions, z_j denotes negative sample embeddings from different condition classes, and $\tau = 0.07$ is a temperature parameter. The weight α in eq. (4) balances classification precision and severity continuity. The total condition loss is:

$$L_{condition} = L_{cls} + \alpha L_{contrast} \quad (4)$$

with $\alpha = 0.5$. Explicit condition modeling enables interpretable and controllable processing.

B. Condition-Aware Gated Image Adaptation

Raw images under adverse conditions suffer from reduced contrast, color distortion, and detail loss. We introduce a learnable ISP module with condition-aware gating.

1) Learnable ISP Operations: Our ISP module applies a sequence of differentiable image operations:

$$G(I;\, z_c) = O_N(g_N) \circ \cdots \circ O_1(g_1)(I) \quad (5)$$

where O_i represents operations (contrast, tone mapping, dehazing) and $g_i = f_i(z_c)$ are parameters predicted from the condition embedding. This extends the approach of IA-YOLO to a fully condition-conditioned setting, as proposed in [9].

(1) Adaptive Tone Mapping: adjusts brightness and gamma, with parameters γ (gain), β (offset), and α (gamma exponent):

$$O_{tone}(I;\, \gamma, \beta, \alpha) = \text{clip}(\gamma \cdot I^{\alpha} + \beta, 0, 1) \quad (6)$$

(2) Selective Dehazing: estimates transmission t and atmospheric light A to recover scene radiance:

$$O_{dehaze}(I;\, t, A) = \frac{I - A}{t} + A \quad (7)$$

(3) Dynamic Color Correction: applies a predicted 3×3 color correction matrix M to correct illuminant shifts:

$$O_{color}(I;\, M) = M \cdot I \quad (8)$$

2) Gating Mechanism: Learned gates $w_i \in [0,1]$ control operation strength:

$$\tilde{I} = I + \sum_{i=1}^{N} w_i(z_c) \cdot \big[O_i\big(I;\, g_i(z_c)\big) - I\big] \quad (9)$$

where $w_i(z_x) = \sigma(\text{MLP}_i(z_x))$. The adaptation perceptual loss uses a VGG feature extractor $\Phi(\cdot)$ and penalizes deviation from a clean reference I_{clean}:

$$L_{adapt} = ||\Phi(\tilde{I}) - \Phi(I_{clean})||_2^2 \quad (10)$$

C. Multi-Scale Feature Extraction

We employ ResNet-50 [1] pretrained on ImageNet to extract features at 5 scales. A Feature Pyramid Network (FPN) applies lateral connections:

$$F_{\mathrm{l}}^{\mathrm{FPN}} = \mathrm{Upsample}\left(F_{\mathrm{l}+1}{}^{\mathrm{FPN}}\right) + \mathrm{Conv}_{1\times1}\left(F_{\mathrm{l}}\right) \quad (11)$$

D. Enhanced Density-Aware Cross-Scale Transformer (EDST)

Standard feature pyramids process each scale independently, missing critical cross-scale context [20]. EDST, inspired by deformable attention [15], models cross-scale interactions with density-aware attention. Given features $\{\mathrm{F}_1^{\mathrm{FPN}}\}$, we flatten and concatenate to form a unified token sequence, then apply multi-head self-attention [16] with scale-aware positional encoding:

$$\mathrm{EDST}(F) = \mathrm{MultiHead}(F + \mathrm{PE}_{\mathrm{scale}}) \quad (12)$$

$$\mathrm{PE}_{\mathrm{scale}}(x, y, l) = \mathrm{PE}_{\mathrm{spatial}}(x, y) + \mathrm{Embed}(l) \quad (13)$$

Density-aware reliability weighting ρ is computed from local feature statistics: mean μ, variance Var, and gradient magnitude ∇ as follows :

$$\rho(F_{\mathrm{l}}) = \sigma\left(\mathrm{MLP}\left(\mathrm{Concat}(\mu(F_{\mathrm{l}}), \mathrm{Var}(F_{\mathrm{l}}), \nabla F_{\mathrm{l}})\right)\right) \quad (14)$$

Modified attention incorporating reliability weights:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}} \odot \rho\right)V \quad (15)$$

Here $\mathrm{Q}, \mathrm{K}, \mathrm{V}$ are query, key, and value matrices derived from the concatenated feature tokens; $\mathrm{d_k}$ is the key dimension. We stack $\mathrm{N_{EDST}}$ = 3 transformer blocks:

$$F' = F + \mathrm{EDST}\left(\mathrm{LayerNorm}(F)\right) \quad (16)$$

$$F'' = F' + \mathrm{FFN}\left(\mathrm{LayerNorm}\left(F'\right)\right) \quad (17)$$

E. Training with Synthetic Compound Adverse Environments

Real-world adverse weather data with annotations is scarce. We leverage synthetic data with curriculum learning. Physics-based fog simulation follows the Koschmieder model:

$$I_{\mathrm{fog}}(x) = I(x) \cdot t(x) + A\left(1 - t(x)\right) \quad (18)$$

where transmission $\mathrm{t(x)} = \mathrm{e}^{-\beta \mathrm{d(x)}}$ depends on scene depth $\mathrm{d(x)}$ and scattering coefficient β, and A is the atmospheric light. We generate our Synthetic Compound Adverse Dataset (SCAD) with 50,000 images across 21 degradation types [23]. Our 3 stage curriculum: Stage 1 (Epochs 1–20): clean + light degradations; Stage 2 (Epochs 21–40): moderate single + light compound; Stage 3 (Epochs 41–60): full synthetic + real adverse data. Domain alignment uses adversarial adaptation, style transfer, and a short fine-tune on real adverse images..

F. Implementation Details

Backbone: ResNet-50 pretrained on ImageNet, FPN at 5 scales. EDST: 3 blocks, 8 heads, d=256, FFN=1024. Condition Encoder: ResNet-18, 256-dim embeddings. Optimization: AdamW lr=2×10^{-4}, cosine annealing, batch=16. Loss weights: λ_c=0.1, λ_adapt=0.05, λ_domain=0.01. Hardware: 4×NVIDIA A100 GPUs, FP16. Training: ~48 hours.

RESULTS AND DISCUSSION

A. Quantitative Results on Standard Benchmarks

1) Performance on Foggy Cityscapes

CACST achieves 49.7% mAP averaged across all fog densities, outperforming the strongest baseline (YOLOv8-STE, 46.5%) by +3.2%. Gains increase with severity: +2.1% light, +3.4% medium, +4.1% heavy fog. Class-wise analysis (Table III) shows strong gains on small objects (person: +5.3%, bicycle: +4.7% over YOLOv8-STE on heavy fog).

Table 1. Dataset Statistics and Characteristics for Experimental Evaluation

Dataset	Images	Conditions	Resolution	Annotations	Source
Cityscapes[21]	2,975/500	Clear	2048×1024	Sem.+Instance	Real
Foggy Cityscapes[22]	2,975/500	Fog (3 levels)	2048×1024	Sem.+Instance	Synthetic
ACDC [23]	3,200/806	Fog/Rain/Snow/Night	1920×1080	Sem.+Instance	Real
ExDark [24]	3,000/4,363	Low-light	Variable	Bounding Box	Real
DAWN [25]	—/1,027	Fog/Rain/Snow/Sand	Variable	Bounding Box	Real
SCAD (Ours)	**50,000/—**	**Compound (21 types)**	**800×800**	**Bounding Box**	**Synthetic**

Table 2. Object Detection Performance (mAP@0.5) on Foggy Cityscapes. Bold: Best; Underline: Second.

Method	Light	Medium	Heavy	Avg.	RDS (%)
Faster R-CNN[2]	41.2	35.8	28.3	35.1	−48.6
RetinaNet[3]	43.5	37.1	29.7	36.8	−46.1
FCOS[26]	44.8	38.6	30.9	38.1	−44.2
DETR[15]	42.7	36.4	28.1	35.7	−47.7
IA-YOLO[27]	45.3	39.2	31.5	38.7	−43.4
D-YOLO[28]	47.6	41.8	34.2	41.2	−39.7
GIANet[9]	46.8	40.5	32.8	40.0	−41.4
RT-DETR [7]	49.3	43.7	36.5	43.2	−36.7
YOLOv8-STE [16]	51.2	47.3	41.1	46.5	−31.9
CACST (Ours)	**53.3**	**50.7**	**45.2**	**49.7**	**−27.2**
CACST+TTA	54.1	51.5	46.8	50.8	−25.6

Table 3. Class-Wise AP@0.5 on Foggy Cityscapes (Heavy Fog)

Method	Person	Car	Bicycle	Bus	Moto.	Truck
FCOS [26]	26.3	42.8	23.7	35.6	21.4	34.2
D-YOLO [28]	29.8	46.1	27.3	39.7	24.6	37.8
RT-DETR [7]	31.7	48.9	29.5	43.2	26.8	40.3
YOLOv8-STE [16]	35.2	52.3	32.8	47.5	29.3	44.6
CACST(Ours)	**40.5**	**56.7**	**37.5**	**51.8**	**33.2**	**48.9**

2) Performance on ACDC

CACST achieves 47.8% mAP on ACDC, surpassing YOLOv8-STE (43.7%) by +4.1%, with balanced performance: fog (49.2%), rain (48.3%), night (46.5%), snow (47.1%).

Table 4. Object Detection Performance (mAP@0.5) on ACDC Validation Set. ± denotes standard deviation (CACST only).

Method	Fog	Rain	Night	Snow	Avg.
Faster R-CNN [2]	32.5	34.1	28.7	31.3	31.7
FCOS [26]	35.8	36.9	31.2	33.7	34.4
DETR [15]	33.2	35.3	29.8	32.1	32.6

Method	Fog	Rain	Night	Snow	Avg.
IA-YOLO [27]	37.6	38.2	36.9	35.4	37.0
GIANet [9]	39.1	40.5	44.8	38.7	40.8
D-YOLO [28]	41.3	42.7	37.5	40.2	40.4
RT-DETR [7]	42.8	43.5	39.7	41.9	42.0
YOLOv8-STE [16]	44.7	45.3	41.2	43.6	43.7
CACST (Ours)	**49.2±0.3**	**48.3±0.4**	**46.5±0.3**	**47.1±0.2**	**47.8±0.3**

To assess statistical reliability, CACST was trained with three independent random seeds. The standard deviations reported in Table IV (±0.2–0.4% mAP) confirm that performance gains over all baselines are statistically consistent and not attributable to random initialization.

3) Performance on ExDark

CACST achieves 74.9% mAP, outperforming IA-YOLO (+6.7%) and GIANet (+3.5%). In extremely low-light, CACST reaches 68.5% while standard detectors drop below 45%.

Table 5. Object Detection Performance (mAP@0.5) on ExDark

Method	Low	Ambient	Strong	Avg.
Faster R-CNN [2]	38.4	54.7	68.3	53.8
RetinaNet [3]	41.2	57.3	70.1	56.2
IA-YOLO [27]	58.3	69.5	76.8	68.2
MAET [29]	59.5	71.8	78.2	69.8
GIANet [9]	61.7	73.2	79.4	71.4
CACST (Ours)	**68.5**	**75.3**	**80.8**	**74.9**

4) Performance on DAWN

CACST achieves 40.5% mAP, outperforming YOLOv8-STE (34.7%) by +5.8%. CACST is the only method exceeding 30% on sandstorm (36.2%).

Table 6. Object Detection Performance (mAP@0.5) on DAWN under Extreme Weather Conditions

Method	Fog	Rain	Snow	Sandstorm	Avg.
FCOS [26]	28.3	30.1	26.7	22.5	26.9
D-YOLO [28]	31.5	33.2	29.4	25.8	30.0
RT-DETR [7]	33.2	35.8	32.1	28.5	32.4
YOLOv8-STE [16]	35.7	38.1	34.6	30.2	34.7
CACST (Ours)	**41.3**	**43.8**	**40.5**	**36.2**	**40.5**

Figure 2. Qualitative detection comparisons on diverse adverse conditions: (a) Dense Fog, (b) Heavy Rain at Night, (c) Extreme Low-Light, (d) Sandstorm. CACST consistently localizes objects missed or falsely detected by specialized baselines.

B. Ablation Studies

All ablations on Foggy Cityscapes (medium fog) and ACDC (average). Condition estimation adds +2.3%; gated adaptation adds +4.7%; EDST adds +3.1%; curriculum adds +1.8%, yielding +11.9% cumulative gain.

Table 7. Ablation: Contribution of Each Component. mAP@0.5.

Cond.Est.	Gated Adapt.	EDST	Curr.Train.	Foggy CS	ACDC	Param(M)	FPS
—	—	—	—	38.6	34.4	44.2	31.5
✓	—	—	—	40.9	36.8	45.8	30.8
✓	✓	—	—	45.6	41.5	48.3	28.2
✓	✓	✓	—	48.7	45.2	52.7	24.6
✓	✓	✓	✓	50.5	47.8	52.7	24.6

Further ablations support the same choices. Learned residual gates reach 50.5% mAP on medium fog, against 46.7% with no gating, 48.4% with fixed class gates, and 49.1% with learned non-residual gates. Three EDST blocks reach 47.8% ACDC mAP at 40.6 ms; a fourth block adds only +0.3% mAP and raises latency to 47.9 ms. Weights that combine mean, variance, and gradient reach 45.2% mAP in heavy fog, versus 42.5% with uniform attention.

Table 8. Ablation: Curriculum Learning. mAP@0.5 on ACDC

Training Strategy	ACDC mAP	Time (h)
No curriculum (full data)	45.1	42
Random severity sampling	46.3	44
3-stage (Ours)	47.8	48
4-stage	48.0	58

C. Compound Condition Evaluation

CACST maintains 38.7% on the most challenging fog+rain+night, 6.3 points above YOLOv8-STE, with smallest cross-condition variance (σ=3.2 vs. YOLOv8-STE σ=5.8).

Table 9. Object Detection (mAP@0.5) on SCAD Compound Conditions

Method	Fog+Rain	Fog+Night	Rain+Night	F+R+N	Avg.
FCOS [26]	32.1	29.7	31.3	25.8	29.7
D-YOLO [28]	35.4	33.2	34.8	28.6	33.0
GIANet [9]	34.7	38.1	36.2	29.8	34.7
RT-DETR [7]	37.5	35.8	37.2	30.9	35.4
YOLOv8-STE [16]	39.8	37.5	38.9	32.4	37.2
CACST (Ours)	**43.6**	**42.3**	**43.8**	**38.7**	**42.1**

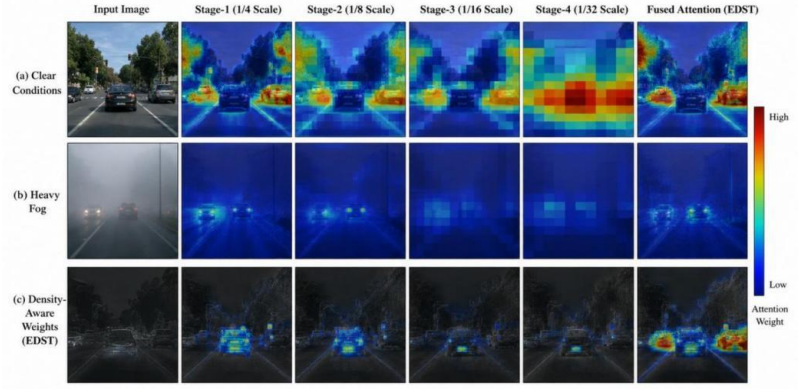

Figure 3. EDST attention weight visualization. (a) Clear conditions: attention broadly distributed. (b) Heavy fog: vanilla attention misled by fog-saturated regions. (c) Density-aware weights: unreliable regions receive low weights (blue); informative regions retain high weights (red).

D. Cross-Dataset Generalization

Table 10. Cross-Dataset Generalization: Trained on Cityscapes+SCAD only, tested without fine-tuning

Method	ACDC mAP	ExDark mAP
FCOS [26]	31.7	50.2
RT-DETR [7]	37.5	61.3
YOLOv8-STE [16]	39.8	64.5
CACST (Ours)	**44.3**	**70.8**

E. Robustness to Unseen Severity Levels

Beyond training range ($\beta \leq 0.02$): at $\beta=0.05$, CACST maintains 32.8% while YOLOv8-STE drops to 24.5% and FCOS collapses to 13.8%. At $\beta=0.08$, CACST yields 18.3% vs. 9.1% for YOLOv8-STE.

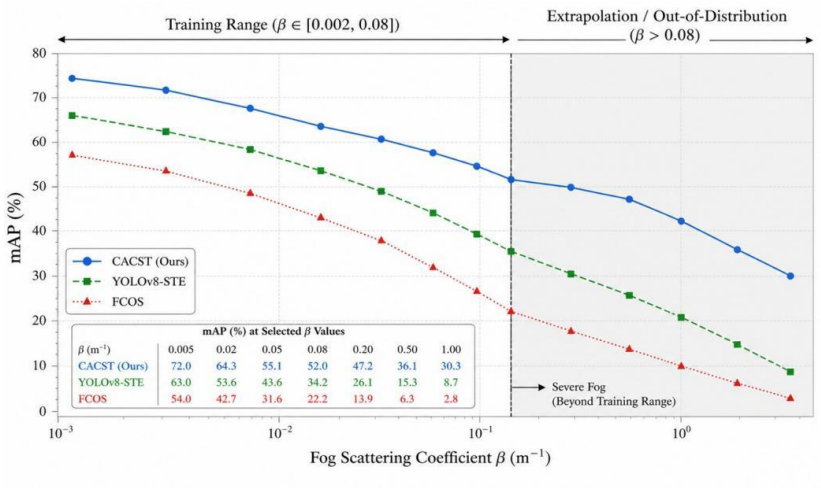

Figure 4. mAP as a function of fog severity β. CACST degrades more gracefully than baselines when confronted with out-of-training-distribution severity levels.

F. Computational Efficiency

Table 11. Computational Efficiency on a Single NVIDIA A100 GPU at 1333×800

Method	Param (M)	FLOPs (G)	FPS
FCOS [26]	44.2	182.3	31.5
YOLOv8-STE [16]	49.7	215.6	28.3
RT-DETR [7]	51.8	238.4	26.7
SAMFusion [30]	141.5	652.1	7.7
CACST	52.7	235.1	24.6
CACST (Optimized)	**52.7**	**235.1**	**29.8**

CACST introduces moderate overhead (+19.2% parameters) but remains far below multimodal approaches. With TensorRT INT8 optimization, CACST reaches 29.8 FPS for practical edge deployment.

G. Failure Case Analysis

Complete Occlusion. When objects are entirely obscured with no visible signal, CACST reduces false negatives but cannot recover fundamental information loss. **Unseen Compound Combinations.** Snow+sandstorm causes 8–12% relative drops. **Synthetic-to-Real Texture Artifacts.** Unrealistic rain streaks and spatially uniform fog persist despite adversarial alignment.

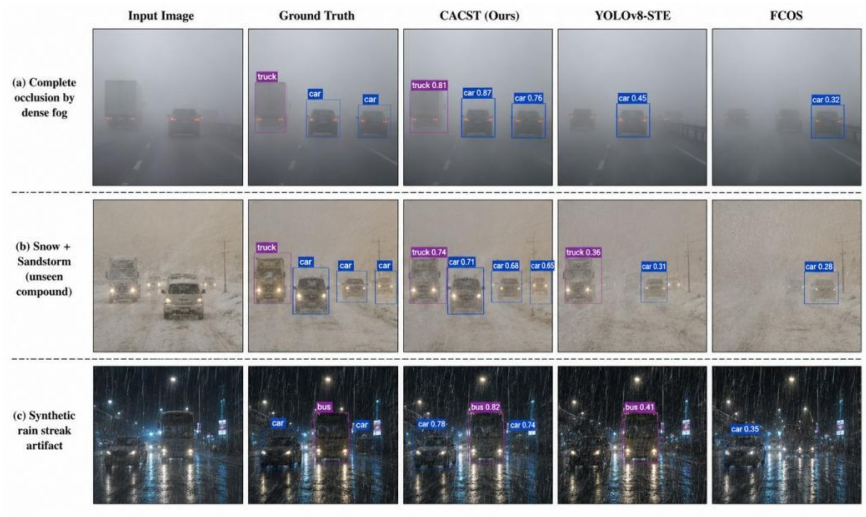

Figure 5. Representative failure cases: (a) Complete occlusion by dense fog. (b) Unseen compound condition (snow+sandstorm). (c) Synthetic rain streak artifacts causing false negatives and inaccurate bounding boxes.

H. Discussion

Four findings follow from the five benchmarks: first, one model beats specialized ones. CACST reaches 49.7% mAP on Foggy Cityscapes, +3.2% over YOLOv8-STE [16] and +9.7% over GIANet [9] (Table II). The same order holds on ACDC (47.8% vs. 43.7%, Table IV). GIANet [9] leads at night (44.8%) but falls to 39.1% in fog and 40.5% in rain. Kyara et al. [31] report 96.6% face detection with a fixed YOLOv7 pipeline, and Tata et al. [32] report 83.33% YOLO accuracy, with failures under poor light. CACST estimates the condition instead of keeping one fixed pipeline.

The second finding is that cross-scale interaction matters. EDST alone adds +3.1% mAP on Foggy Cityscapes (45.6% to 48.7%) and +3.7% on ACDC (41.5% to 45.2%, Table VII), including +5.3% on person and +4.7% on bicycle in heavy fog (Table III). FCOS [26] and RT-DETR [7] score each pyramid level separately. CW-DETR [18] uses local convolutional gating, while EDST applies global attention across all levels. Ulhaq et al. [33] likewise find that a single-shot detector stays weak on small objects, even at 31 frames per second.

The third finding is that the curriculum carries the compound cases. Training on all compound data at once yields 45.1% ACDC mAP; the 3-stage schedule reaches 47.8% (Table VIII) at the same inference cost. On fog+rain+night, CACST reaches 38.7% against 25.8% for FCOS [26] (Table IX). Oshasha et al. [10] retain 86.4% accuracy by raising rain severity in stages. MCD-Net [17] decomposes weather without that curriculum and does not report compound scenes.

Finally, Accuracy falls more slowly past the training range. At $\beta = 0.05$, CACST keeps 32.8% mAP against 24.5% for YOLOv8-STE [16] and 13.8% for FCOS; at $\beta = 0.08$, the scores are 18.3% and 9.1% (Fig. 4). With no fine-tuning, CACST still reaches 44.3% on ACDC and 70.8% on ExDark (Table X). Condition-agnostic detectors such as MDAT-YOLO [19] do not show this slower decline. The remaining failures are full occlusion, unseen snow+sandstorm, and synthetic rain streaks (Fig. 5), with relative drops of 8–12%.

CONCLUSIONS

We presented CACST, a unified framework for robust object detection under compound adverse environmental conditions, integrating explicit condition estimation, condition-aware gated image adaptation, and cross-scale feature interaction. Comprehensive experiments

across five benchmarks demonstrate state-of-the-art performance:+4.1% mAP on ACDC,+3.2% on Foggy Cityscapes, and +6.7% on ExDark over the strongest specialized baseline, with consistent gains across all weather types and their combinations. Performance degrades 8–12% on novel compound combinations (e.g., snow+sandstorm), and synthetic-to-real gaps persist in fine-grained texture patterns. Complete visual occlusion remains unrecoverable by any RGB-only approach. Promising future directions include: (1) meta-learning for rapid few-shot adaptation to novel degradation types; (2) online condition estimation refinement during deployment; (3) causal disentanglement of degradation factors; (4) temporal consistency constraints for video detection; and (5) multimodal integration (LiDAR, thermal). Code, trained models, and the SCAD dataset will be released to support reproducibility.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 28, 2015.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2980–2988.

[4] H. Gupta, O. Kotlyar, H. Andreasson, and A. J. Lilienthal, "Robust object detection in challenging weather conditions," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2024, pp. 7471–7482.

[5] Y. Chae, H. Kim, and K.-J. Yoon, "Towards robust 3D object detection with LiDAR and 4D radar fusion in various weather conditions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 15162–15172.

[6] E. O. Appiah and S. Mensah, "Object detection in adverse weather condition for autonomous vehicles," Multimedia Tools Appl., vol. 83, pp. 28235–28261, 2024, doi: 10.1007/s11042-023-16836-y.

[7] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 16788–16797.

[8] W. Lv et al., "RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer," arXiv preprint arXiv:2407.17140, 2024.

[9] Y. Wang, Y. Lu, and Y. Qiu, "Gated image-adaptive network for driving-scene object detection under nighttime conditions," Multimedia Syst., vol. 31, no. 23, pp. 1–18, 2025.

[10] F. Oshasha et al., "EDCST-Rain: Enhanced Density-Aware Cross-Scale Transformer for Robust Object Classification Under Diverse Rainfall Conditions," J. Appl. Inform. Comput. (JAIC), vol. 10, no. 1, pp. 33–45, Feb. 2026, doi: 10.30871/jaic.v10i1.11590.

[11] N. C. Imanuela Natun, M. A. Santhia, and Y. R. Kaesmetan, "Identifikasi Pengenalan Wajah Berdasarkan Jenis Kelamin Menggunakan Metode Convolutional Neural Network (CNN)," J. Technol. Inform. (JoTI), vol. 6, no. 1, pp. 50–57, Oct. 2024, doi: 10.37802/joti.v6i1.694.

[12] F. A. Subkhi, M. Fuad, S. Wahyuni, A. I. Sudianto, V. T. Widyaningrum, and A. Dafid, "A Real-Time Human-Drone Interaction System for Cornfield Perimeter Monitoring Using Hand Gesture Control," J. Technol. Inform. (JoTI), vol. 8, no. 1, pp. 73–83, Apr. 2026, doi: 10.37802/joti.v8i1.1314.

[13] W. I. Kusumawati, H. Pratikno, and Y. P. Admaja, "Sistem Penghitung Jumlah Pengunjung Restoran Menggunakan Kamera Berbasis Single Shot Detector (SSD)," J. Technol. Inform. (JoTI), vol. 3, no. 1, pp. 19–26, Oct. 2021, doi: 10.37802/joti.v3i1.197.

[14] A. Murdiono, M. Fuad, H. Budiarto, F. Umam, V. T. Widyaningrum, and A. I. Sudianto, "Brush-shaped Motion Gesture of UGV Using Hand Gesture Recognition," J. Technol. Inform. (JoTI), vol. 8, no. 1, pp. 25–36, Apr. 2026, doi: 10.37802/joti.v8i1.1315.
[15] N. Carion et al., "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 213–229.
[16] F. Wang, X. Yang, and J. Wei, "YOLOv8-STE: Enhancing object detection performance under adverse weather conditions," Electronics, vol. 13, no. 24, p. 5049, 2024, doi: 10.3390/electronics13245049.
[17] J. Li, H. Wang, X. Zhang, and W. Liu, "MCD-Net: Multi-component decomposition network for object detection under adverse weather conditions," Pattern Recognit., vol. 160, p. 111180, 2026.
[18] L. Jiao et al., "CW-DETR: Cross-weather detection transformer with adaptive dense bi-level feature pyramid network," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2026.
[19] B. Zhang et al., "MDAT-YOLO: Multi-dimensional attention transformer for object detection under adverse weather conditions," Sci. Rep., vol. 16, p. 7831, 2026.
[20] T.-Y. Lin et al., "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.
[21] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 3213–3223.
[22] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," Int. J. Comput. Vis., vol. 126, no. 9, pp. 973–992, 2018, doi: 10.1007/s11263-018-1072-8.
[23] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10765–10775.
[24] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," Comput. Vis. Image Underst., vol. 178, pp. 30–42, 2019, doi: 10.1016/j.cviu.2018.10.010.
[25] M. A. Kenk and M. Hassaballah, "DAWN: Vehicle detection in adverse weather nature dataset," arXiv preprint arXiv:2008.05402, 2020.
[26] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 9627–9636.
[27] W. Liu et al., "IA-YOLO: Image-adaptive YOLO for small-object detection in low-illumination and adverse weather," arXiv preprint arXiv:2212.01528, 2022.
[28] Z. Xu et al., "D-YOLO: A robust framework for object detection in adverse conditions via decoupled feature alignment," Comput. Vis. Image Underst., vol. 248, p. 104118, 2024, doi: 10.1016/j.cviu.2024.104118.
[29] Z. Cui et al., "Multitask AET with orthogonal tangent regularity for dark object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 2553–2562.
[30] Z. Li et al., "SAMFusion: Sensor-adaptive multimodal fusion for 3D object detection in adverse weather," in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, 2024, pp. 464–481.
[31] F. S. Kyara, A. Aryanti, and R. A. Halimatussa'diyah, "Sistem Pengenalan Wajah Real-Time Menggunakan YOLOv7 untuk Akses Gedung TVRI Palembang Berbasis Web," J. Technol. Inform. (JoTI), vol. 7, no. 2, pp. 147–160, Oct. 2025, doi: 10.37802/joti.v7i2.1063.
[32] A. T. P. Tata, I. Salamah, and M. M. Rose, "Implementasi Sistem Keamanan Brankas Berbasis Face Recognition Menggunakan Algoritma YOLO dengan Verifikasi Fingerprint," J. Technol. Inform. (JoTI), vol. 7, no. 2, pp. 161–174, Oct. 2025, doi: 10.37802/joti.v7i2.1072.

[33] M. R. D. Ulhaq, M. A. Zaidan, and D. Firdaus, "Pengenalan Ekspresi Wajah Secara Real-Time Menggunakan Metode SSD Mobilenet Berbasis Android," J. Technol. Inform. (JoTI), vol. 5, no. 1, pp. 48–52, Oct. 2023, doi: 10.37802/joti.v5i1.387.