Article

Not peer-reviewed version

A Lightweight Degradation-Aware Framework for Robust Object Detection in Adverse Weather

Seungun Park , Jiakang Kuai , Hyunsu Kim , Hyunseong Ko , ChanSung Jung * , Yunsik Son *

Posted Date: 26 November 2025

doi: 10.20944/preprints202511.2035.v1

Keywords: adverse weather object detection; degradation-aware detection; image enhancement for detection; lightweight deep learning; boundary refinement; semantic feature refinement; differentiable image processing

 doi:10.20944/preprints202511.2035.v1

Article

A Lightweight Degradation-Aware Framework for Robust Object Detection in Adverse Weather

Seungun Park [1], Jiakang Kuai [2], Hyunsu Kim [1], Hyunseong Ko [1], ChanSung Jung [3,*] and Yunsik Son [1,*]

[1] Department of Computer Science and Artificial Intelligence, Dongguk University, Seoul 04620, Republic of Korea
[2] Department of Computer Science and Engineering, Dongguk University, Seoul 04620, Republic of Korea
[3] Department of Game Contents, Wonkwang University, Iksan City, Jeonbuk 54538, Republic of Korea
* Correspondence: placomp@wku.ac.kr (C.J.); sonbug@dongguk.edu (Y.S.)

Abstract

Object detection in adverse weather remains challenging due to the simultaneous degradation of visibility, structural boundaries, and semantic consistency. Existing restoration-driven or multi-branch detection approaches often fail to recover task-relevant features or introduce substantial computational overhead. To address this problem, DLC-SSD, a lightweight degradation-aware framework for detecting robust objects in bad weather environments, is proposed. The framework is based on an integrated restoration and refining strategy that performs image-level degradation correction, structural information enhancement, and semantic expression refinement in stages. First, the Differentiable Image Processing (DIP) module performs low-cost enhancement to adapt to global and local degradation patterns. After that, the Lightweight Edge-Guided Attention (LEGA) module uses a fixed Laplacian-based structural dictionary to reinforce the boundary cues in shallow, high-resolution features. Finally, the Content-aware Spatial Transformer with Gating (CSTG) module captures long-distance contextual relationships and refines the deep semantic representation by suppressing noise. These components are jointly optimized end-to-end with the single shot multibox detection (SSD) backbone. In rain, fog, and low-light conditions, DLC-SSD demonstrated more stable performance than conventional detectors and maintained a quasi-real-time inference speed, confirming its practicality in intelligent monitoring and autonomous driving environments.

Keywords: adverse weather object detection; degradation-aware detection; image enhancement for detection; lightweight deep learning; boundary refinement; semantic feature refinement; differentiable image processing

1. Introduction

With the growing demand for intelligent urban management, video surveillance has become a key tool in monitoring road conditions, ensuring traffic safety, and supporting incident response [1]. Object detection plays a key role in such a real-world system; however, maintaining stable performance under adverse weather conditions, such as rain, fog, low illumination, and night shooting, is challenging because the image quality is severely degraded. Under these conditions, blurring, contrast reduction, particulate noise, and structural distortion caused by scattering occur at the same time, blurring the appearance information of the object and blurring the boundary with the background [2,3]. In particular, since high accuracy, real-time processing power, and lightweight are essential in the actual deployment environment, existing heavy restoration networks or high-cost detection models have practical limitations.

Various approaches have been proposed to address the problem of bad-weather object detection, but fundamental limitations remain. Image restoration-based methods often focus on improving human visual quality, so they cannot sufficiently restore structural and semantic clues required by

 doi:10.20944/preprints202511.2035.v1

object detectors [4–6]. Recent studies in other imaging domains have also shown that visually pleasing restoration or reconstruction does not necessarily lead to improved task performance, underscoring the importance of task-driven enhancement strategies tailored to downstream detection objectives [7,8]. On the other hand, adding a separate auxiliary decoder or a multi-branch inside the detector is challenging for real-time performance due to increased parameter count and computational cost. In addition, deep CNN-based detectors lose high-frequency information such as boundary and texture during the repetitive downsampling process, and this structural loss is further intensified in bad weather conditions, making them vulnerable to small objects or objects with blurred boundaries [9–13].

To overcome these limitations, this study proposes a lightweight, unified refinement pipeline that corrects the image-level, structural, and semantic-level degradation in a stepwise manner. The first step minimizes global and low-frequency degradation of the input image at minimum cost; the second step strengthens the boundary cues in shallow, high-resolution features; and the last step performs global context-based semantic alignment in deep features. These three processes do not operate independently of each other. However, they are designed in a complementary manner around detection performance, which simultaneously improves the robustness and expressiveness of single shot multibox detector (SSD)-based object detectors [14,15].

The main contributions of this paper are summarized as follows:

- A lightweight, multi-filter task-driven differentiable image processing (DIP) module is introduced to mitigate the mismatch between image restoration and object detection, enabling adaptive enhancement tailored to diverse degradation patterns.
- A Lightweight Edge-Guided Attention (LEGA) mechanism is designed to reinforce structural cues in shallow high-resolution feature maps using a fixed Laplacian prior. This module improves boundary representation without introducing additional learnable parameters.
- A Content-aware Spatial Transformer with Gating (CSTG) is proposed to jointly strengthen global contextual reasoning and local semantic selectivity through a compact architecture. Integrated seamlessly with the multi-scale feature hierarchy of SSD, CSTG enhances semantic separation for small or blurred objects.
- A unified hierarchical degradation-aware pipeline is constructed by integrating DIP, LEGA, and CSTG, achieving robust performance across rain, fog, and low-light conditions while maintaining near real-time efficiency.

2. Related Work

2.1. Adverse Weather Object Detection

Object detection in adverse weather environments has been extensively studied through various structural variations. Networks with two-pronged or multi-branch structures, such as DSNet and D-YOLO, aim to enhance robustness by incorporating recovery subnetworks or by fusing blurred and sharp features [11,12]. AK-Net decomposes weather degradation with multiple sub-degradation factors such as rain, fog, and water droplets to improve the performance of small object detection in complex environments [13]. Transformer-based techniques, including WRRT-DETR, leverage multi-head self-attention to enhance long-range contextual modeling and semantic discrimination capabilities [16,17]. On the other hand, some techniques, such as ClearSight, adopt a preprocessing-oriented strategy, applying a deep enhancement module and then inputting images into the detector [18]. Representative approaches and their properties are summarized in Table 1.

 doi:10.20944/preprints202511.2035.v1

Table 1. Representative methods for object detection under adverse weather conditions.

Study	Key Idea	Limitations
D-YOLO [12]	Hazy/clear feature fusion via attention	Heavy fusion; auxiliary subnetworks; no image-level enhancement
DSNet [11]	Joint visibility enhancement + detection	Difficult loss balancing; fog-specific; high complexity
AK-Net [13]	Weather degradation separation and feature fusion	Multiple subnetworks; limited generalization across conditions
WRRT-DETR [16]	Global self-attention for long-range context	High computational cost; unsuitable for real-time
ClearSight [18]	Deep dehazing before detection	Not jointly optimized; expensive preprocessing

Despite meaningful progress, these approaches have common limitations. Restoration-detection hybrid structures typically require heavy components such as decoders, auxiliary subnetworks, and multipath branches, which significantly increase the number of parameters and computational cost. Models specific to a particular type of degradation generalize poorly to other conditions, and Transformer-based designs incur high memory usage and slow inference speeds, making real-time applications challenging. Preprocessing-based enhancement techniques also exhibit weak task alignment with real-world detection objectives, as the enhancement phase is optimized independently of the detection objectives.

2.2. Differentiable Image Processing and Task-Driven Enhancement

Image-level enhancement has been actively studied to mitigate visibility degradation caused by factors such as rain, fog, and low illumination. Techniques such as ZeroDCE focus on correcting illuminance imbalances and contrast degradation without the need for paired supervised data, demonstrating clear improvements in visual quality [6]. However, since these enhancement-only designs operate independently of detection tasks, they show limitations in providing performance improvements in complex adverse weather environments. Detector-based enhancement techniques such as IA-YOLO, GDIP, and ERUP-YOLO incorporate enhancement modules into the detection backbone, but in many cases depend on decoder-type structures or parameter-rich filtering operations to reduce real-time efficiency [19–21]. Furthermore, augmentation-centered approaches also reveal that image-level correction alone does not consistently translate into improvements on detection metrics under real adverse-weather conditions [22]. Moreover, pixel-level enhancements alone do not sufficiently restore the structural cues or semantic consistency required for stable detection. Representative enhancement-based techniques and their limitations are summarized in Table 2.

 doi:10.20944/preprints202511.2035.v1

Table 2. Representative methods based on differentiable image processing and task-driven enhancement.

Study	Key Idea	Limitations
ZeroDCE [6]	Zero-reference curve estimation; lightweight pixel-wise correction	Not task-driven; limited generalization to rain/fog; no multi-scale feature modeling
IA-YOLO [19]	Adaptive enhancement via integrated module	Add-on enhancement increases overhead; limited filter diversity
GDIP [20]	Parameterized image filters embedded into detection pipeline	Heavy operations on full-resolution images; limited structural modeling
ERUP-YOLO [21]	Unified image-adaptive filtering with Bezier-based pixel and kernel operations	Limited robustness to over-enhancement; no feature-level weather modeling
Augment + YOLOv5 [22]	Training with real all-weather data and physics/GAN-based augmentation	Synthetic noise often unstable; no joint enhancemnet-detection optimization

Task-driven enhancements remain an unsolved problem because existing techniques either operate separately from detection pipelines or incur significant computational costs. The DIP module design proposed in this study resolves this gap by introducing a fully differentiable structure and a lightweight filter-parameter prediction mechanism based on a small CNN. Multi-filter combinations involving noise cancellation, sharpening, and pixel-wise correction are implemented as differentiable operations without relying on high-resolution reconstruction networks. This allows the enhancement process to be co-optimized directly with SSD detection loss and naturally coupled with subsequent structural and semantic refinement modules such as LEGA and CSTG.

2.3. Edge-Aware and Laplacian-Based Structural Refinement

The edges and contours, which are structural clues, are easily damaged in rain, fog, and low illumination environments. Since the deep CNN backbone gradually loses high-frequency information through iterative downsampling and convolution, the degradation of structural features directly affects the accuracy of small-object detection and position estimation. Previous studies have combined Laplacian filters, edge pyramids, and contour recognition mechanisms to complement structural information. However, these approaches often require multi-scale reconstruction, decoder networks, and additional learning-based gradient extraction, which increases computational cost and limits their use in lightweight detection frameworks. Table 3 summarizes representative edge-aware structural refinement methods.

Table 3. Representative methods employing edge-aware and Laplacian-based structural refinement.

Study	Key Idea	Limitations
Laplacian Pyramid Reconstruction [23]	Refines semantic features using Laplacian pyramid–based boundary reconstruction	Requires multi-level decoding; high computational overhead
B2Net [24]	Explicit boundary extraction and fusion for camouflaged objects	Learnable boundary extractor increases parameters; less suitable for real-time
MEGANet [25]	Multi-scale edge features guide attention for weak-boundary segmentation	Multi-branch edge pathways raise complexity and memory usage
BorderDet [26]	Enhances dense detectors by modeling explicit border cues	Additional border heads and regression terms increase inference cost

 doi:10.20944/preprints202511.2035.v1

Many edge-based techniques have been primarily applied in image restoration and have not been fully utilized in the detection backbone. As a result, it does not take full advantage of the opportunity to enforce structural refinement in shallow, high-resolution feature layers, where object boundaries are best preserved. To overcome these limitations, the LEGA module combines parameter-free Laplacian kernels with small-scale gating mechanisms, enabling effective boundary enhancement at a minimal computational cost. The design highlights the structural elements of the degradation layer, complementing the image-unit correction of DIP and the semantic-unit purification of CSTG.

2.4. Transformer and Gated Attention for Weather-Adaptation

Self-attention mechanism was recently introduced to enhance feature representation in complex weather environments. Methods such as Weather-aware RT-DETR, and YOLO-DH leverage multi-head attention modules, Transformer encoders, and gating-based fusion to rearrange fog or noise-damaged deep features [16,27,28]. These designs are highly expressive, improving long-range dependency modeling and semantic discrimination skills.

However, Transformer-based architectures inherently require substantial memory and computational resources. Multilayer encoders, channel expansion, and multi-head attention stacks are not suitable for real-time or edge environments. Furthermore, focusing on the entire backbone or feature pyramid can lead to redundant computations, especially in one-stage detectors where shallow and deep layers play different roles. Table 4 presents a typical Transformer/gating-based approach.

Table 4. Representative Transformer- and gated-attention-based methods for weather-adaptive detection.

Study	Key Idea	Limitations
Weather-aware RT-DETR [27]	Fog-adaptive dual-stream attention for RT-DETR	Inconsistent gains; limited generalization beyond fog
YOLO-DH [28]	Wavelet-guided dehazing and adaptive attention fusion	Added complexity; enhancement not jointly optimized with detection

The CSTG module introduced in this work mitigates these constraints by confining Transformer operations to the deep extra layers of SSD and by integrating a lightweight content-aware gating mechanism. This selective refinement strategy maintains the advantages of global context modeling while reducing overhead, enabling effective semantic enhancement in adverse weather conditions.

2.5. Summary of Research Gap

Adverse weather detection has been approached at the level of image enhancement, structural refinement, and semantic modeling. However, most existing studies have limitations, focusing only on one of these degradation layers and on semantic modeling. However, most existing studies have limitations: they focus on only one of these degradation layers, incur significant computational costs, or fail to maintain consistency between the enhancement and detection phases. Prior enhancement methods are not co-optimized with detectors; Transformer-based refinement techniques are too heavy to be applied to real-time systems; and edge-based techniques rely on multi-scale reconstruction, which impedes lightweight design.

A unified degradation-aware design that simultaneously addresses image-, structure-, and semantic-level deterioration in a lightweight, SSD-compatible manner remains largely underexplored. The integration of DIP, LEGA, and CSTG into a single pipeline provides a compact yet comprehensive solution, offering strong task alignment, computational efficiency, and robustness across various adverse-weather conditions.

 doi:10.20944/preprints202511.2035.v1

3. Proposed Method

3.1. Overall Pipeline Architecture

The proposed framework adopts a unified degradation-aware refinement process that restores image, structural, and semantic information affected by adverse weather. An overview of the complete architecture is shown in Figure 1. The primary processing flow is summarized as follows. This unified combination of DIP, LEGA, and CSTG, along with an SSD, forms the proposed DLC-SSD framework.

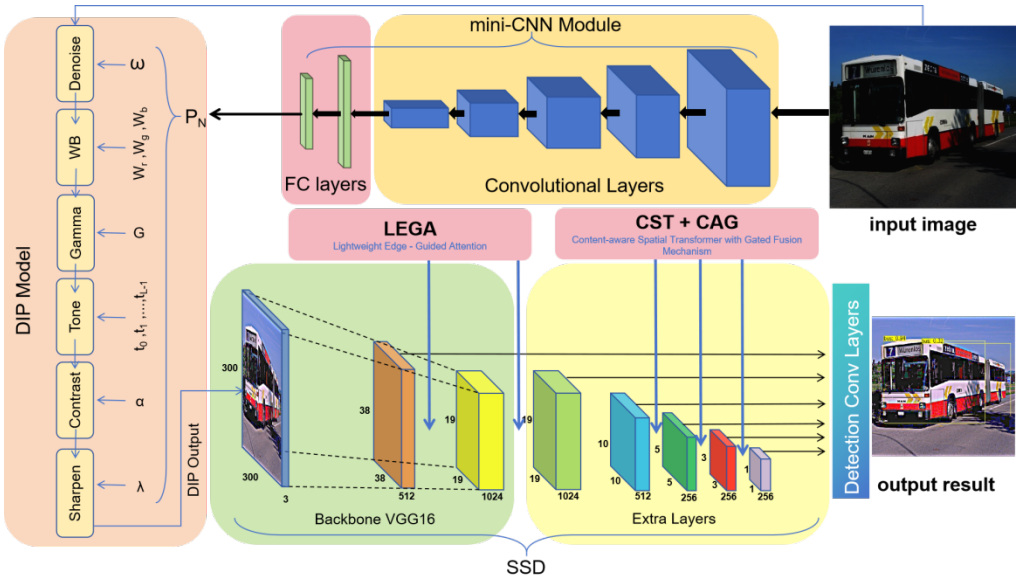

Figure 1. Overall architecture of the proposed DLC-SSD framework.

Given an input image, a lightweight mini-CNN first analyzes its global and local characteristics to predict the parameters required by the DIP module. These predicted parameters control a sequence of learnable and differentiable filters, including sharpening, contrast adjustment, tone and gamma correction, white-balance normalization, and denoising, allowing the system to generate a task-driven enhanced image that is directly optimized for downstream detection.

The enhanced image is then processed by a VGG16-based SSD backbone that extracts multi-scale feature maps [29]. At the shallow stage, the LEGA module injects a fixed Laplacian-based structural prior to strengthen boundaries of weakened objects. This improves early feature stability in regions affected by blur or low visibility. At deeper stages, the CSTG module performs semantic refinement by jointly capturing long-range contextual relations and filtering out noise through content-dependent gating. The resulting refined multi-scale features are fed into SSD detection heads for reliable classification and localization under challenging weather conditions.

3.2. Hierarchical Image-Structure-Semantic Refinement

Adverse weather conditions degrade visual information at multiple levels. To address this, the proposed framework applies refinement hierarchically across three complementary stages: appearance-level enhancement, boundary-level structural reinforcement, and semantic-level contextual refinement. Each stage focuses on a distinct type of degraded information, enabling the overall system to adaptively restore relevant cues while maintaining computational efficiency and stable feature progression throughout the detection pipeline.

3.2.1. Differentiable Image Processing

The DIP stage is designed to restore the visual quality of the input image, which has deteriorated in a bad-weather environment, from an early stage. In this process, DIP dynamically determines the intensity of enhancement optimized for the conditions of each image by analyzing global characteristics,

 doi:10.20944/preprints202511.2035.v1

such as brightness, color balance, tone, and noise density. This adaptive initial restoration is a key factor in ensuring the stability of subsequent structural and semantic purification steps.

DIP operates based on the hyperparameters predicted by the lightweight mini-CNN and has a structure in which filter parameters are trained on low-resolution images (256×256) and applied at the original resolution to reduce the cost of high-resolution processing. Mini-CNN is designed to efficiently extract only the global characteristics of the scene by consisting of a multi-stage convolutional block and a fully connected layer, and it has only 156K parameters, making it suitable for real-time processing. This design enables DIP to perform content-aware enhancements that reflect scene characteristics beyond simple input correction.

The DIP module consists of six differentiable filters, consisting of Denoise, White Balance, Gamma, Contrast, Tone, and Sharpen, all of which are directly optimized through the backpropagation of the network. These filters operate in combination to perform adaptive enhancement tailored to the deterioration pattern of each image, and all mappings are designed to be differentiated, enabling joint optimization based on a single detection loss with the entire detection network. In particular, unlike fixed and task-agnostic enhancement techniques, this DIP module is learnable, detection-aware, and optimized end-to-end via backpropagation. These features make DIP function as a task-driven image enhancement module, closely integrated with the entire detection pipeline beyond simple preprocessing.

Denoise Filter.

The proposed Denoise filter is designed to effectively eliminate wet noise, scattering, and blurring occurring in bad weather by reconstructing the DCP-based restoration technique in a differential form. This filter is based on an atmospheric scattering model, and the input noise image $I(x)$ is expressed as follows:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

Here, $J(x)$ denotes the clean image to be restored, A represents a global atmospheric light, and $t(x)$ is a transmission map. The transmission map is defined as follows based on the scene depth $d(x)$ and the atmospheric scattering coefficient β:

$$t(x) = e^{-\beta d(x)} \tag{2}$$

To restore a clean image $J(x)$, it is essential to estimate A and $t(x)$. To this end, the dark channel of the input image is computed, and then A is estimated as the average value of the corresponding region by selecting the top 1000 brightest pixels. Thereafter, the DCP-based transmission map estimation equation is as follows:

$$t(x) = 1 - \min_{c} \left(\min_{y \in \Omega(x)} \frac{I^c(y)}{A^c} \right) \tag{3}$$

Here, c denotes a color channel, and $\Omega(x)$ represents a local window. In this study, a learnable parameter ω was introduced to control the degree of suppression of the transmission map and generalized as follows:

$$t(x) = 1 - \omega \min_{c} \left(\min_{y \in \Omega(x)} \frac{I^c(y)}{A^c} \right) \tag{4}$$

ω is optimized via backpropagation and enables more robust restoration across various deterioration conditions, such as wet, low-illumination, rain, and fog environments. Since all the above equations are differentiable, the Denoise filter can be trained end-to-end with the entire detection network, enabling detection-aware restoration.

 doi:10.20944/preprints202511.2035.v1

Pixel-wise Filters.

The pixel-wise filters consist of a continuous mapping function that acts directly on the input pixel $P_i = (r_i, g_i, b_i)$. It is the most basic and computationally efficient correction operation in DIP. This filter group consists of four types: White Balance, Gamma, Contrast, and Tone, and all parameters are determined by the values predicted by mini-CNN. Since each operation has an independent pixel-wise conversion structure, the amount of computation is small even in a high-resolution image, and all functions are fully differentiated for input and parameter, so end-to-end learning through detection loss is possible.

The White Balance filter adjusts channel-wise color distortions by applying learnable scaling factors to each RGB component. For an input pixel P_i, the corrected output is obtained through a simple linear transformation,

$$WB(P_i) = (W_r r_i, W_g g_i, W_b b_i) \tag{5}$$

where W_r, W_g, and W_b are the per-channel weighting coefficients predicted by the mini-CNN. This operation provides a stable and differentiable mechanism for balancing color cast under adverse weather conditions.

The Gamma filter modifies global luminance by applying a nonlinear power mapping. For each channel, the output intensity is computed as:

$$G(P_i) = P_i^{\gamma} \tag{6}$$

with γ being a learnable gamma coefficient. This enables the model to reshape the brightness distribution of the input image and to emphasize darker or brighter regions depending on the scene illumination.

To enhance contrast, the Contrast filter interpolates between the original pixel value and a nonlinearly enhanced representation $En(P_i)$:

$$C(P_i) = \alpha \cdot En(P_i) + (1 - \alpha) \cdot P_i \tag{7}$$

The enhanced representation is derived from the pixel's luminance, defined as:

$$Lum(P_i) = 0.27 r_i + 0.67 g_i + 0.06 b_i \tag{8}$$

which is then passed through a smooth cosine-based nonlinear transform,

$$EnLum(P_i) = \frac{1}{2}(1 - \cos(\pi \times (Lum(P_i)))) \tag{9}$$

and finally projected back to the RGB channels through,

$$En(P_i) = P_i \times \frac{EnLum(P_i)}{Lum(P_i)} \tag{10}$$

This formulation allows the contrast filter to enhance intensity variations while maintaining continuous gradients for stable optimization.

Finally, the Tone filter adjusts tonal characteristics using a learnable piecewise polynomial mapping. With tone-curve parameters $t_0, t_1, ..., t_{L-1}$ predicted by the mini-CNN, the output is computed as:

$$T(P_i) = \frac{1}{T_L} \sum_{j=0}^{L-1} clip(L \cdot P_i - j, 0, 1) t_j \tag{11}$$

where $clip(x, 0, 1)$ is an operation for limiting an input value to between 0 and 1. It is defined as follows:

 doi:10.20944/preprints202511.2035.v1

$$clip(x, 0, 1) = min(max(x, 0), 1) \tag{12}$$

This operation is not just clamping; it also serves as a soft weight for each section of the tone curve. That is, when $L \cdot P_i - j$ is inside a specific section, it linearly increases from 0 to 1, determines the contribution to the tone coefficient t_j of the section, and is saturated with 0 or 1 outside the section to create a smooth transition with the adjacent section. This structure configures tone mapping with continuous and section-specific characteristics. Since all operations are differentiable with respect to both input and learning parameters, the entire DIP can be optimized end-to-end via a detection loss.

Sharpen Filter.

The Sharpen filter is inspired by the unsharp masking technique and serves to clearly restore the boundary and microstructure of the object by emphasizing its high-frequency components. The Sharpen operation is defined as the following continuous mapping function:

$$F(x, \lambda) = I(x) + \lambda(I(x) - Gau(I(x))) \tag{13}$$

Here, $I(x)$ is an input image, $Gau(I(x))$ is a Gaussian-blurred image at the exact location, and $\lambda > 0$ is a trainable coefficient for controlling sharpening intensity. Since Gaussian blur extracts low-frequency components, $I(x) - Gau(I(x))$ extracts high-frequency residuals from inputs, λ determines how much to emphasize these residuals.

Since this mapping is completely differentiable with respect to both the input x and the scale factor λ, the degree of sharpening is automatically adjusted during end-to-end optimization of the entire DIP with a detection loss. This can strengthen the blurred object boundaries in challenging weather conditions and provide explicit feature representations that subsequent structural and semantic refinement steps can leverage.

3.2.2. Lightweight Edgie-Guided Attention

In adverse weather conditions, such as rain, fog, and low-light night, the contour and structural cues of the object are blurred, and background noise increases, making the boundary information in the feature map prone to loss. To reduce this structural ambiguity and preserve object shape cues, this study introduces a lightweight structure enhancement module, LEGA. LEGA aims to maintain precise contours and boundaries even when input quality deteriorates. It works in conjunction with CSTG, which performs semantic refinement, to form an image-structured-semantic enhancement flow. In particular, LEGA significantly improves detection stability for small objects or targets with blurred boundaries by directly reinforcing structural information without adding trainable parameters.

LEGA first performs depthwise convolution on the input feature map using the non-learnable fixed Laplacian kernel presented in Figure 2. The Laplacian kernel is a classical boundary detection filter that emphasizes the high-frequency components of the central pixel relative to its surrounding pixels, thereby reliably capturing structural discontinuities, such as edges, corners, and textures, in the image. This creates an edge map E for each channel, complementing the low-level structural information that tends to degrade during downsampling.

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Figure 2. Non-learnable fixed Laplacian kernel for structure-aware edge extraction in LEGA.

The extracted edge map is converted into a structure-based attention mask A by passing through the 1×1 convolution and the sigmoid activation function. This mask highlights structurally important

 doi:10.20944/preprints202511.2035.v1

regions and suppresses background noise and low-frequency components. This process can be expressed in an equation as follows:

$$E = LaplacianConv(F), \qquad A = \sigma(Conv(E)) \tag{14}$$

Here, the $F \in \mathbb{R}^{C\times H\times W}$ is an input feature map, E is a Laplacian-based edge map, and $A \in (0,1)^{C\times H\times W}$ is an attention mask weighted according to structural importance. Finally, the enhanced output feature map F' is calculated as element-wise multiplication as follows:

$$F' = F \odot A \tag{15}$$

where $\odot$ means a multiplication operation by position.

LEGA is applied to high-resolution shallow feature maps located near the input end of the network, such as the conv4_3 and fc7 layers. Since these early-stage feature maps are the stage before structural information is lost through downsampling, they are relatively rich in boundary and outline information and are particularly effective for small objects or targets with blurred boundaries. In addition, LEGA is composed of only a fixed kernel-based depthwise convolution and a shallow 1×1 convolution, so it does not add any learning parameters, and the increase in computation is minimal. This design enables LEGA to efficiently reinforce structural cues computationally, improving the clarity of object boundaries under challenging weather conditions and enhancing the robustness of the overall detection pipeline.

3.2.3. Content-aware Spatial Transformer with Gating

In bad weather conditions, the boundaries of objects are blurred by rain, fog, and low illumination, background noise increases, and it is difficult to capture such complex deformations with only local convolution-based representations. In particular, the fixed receptive field of CNNs has inherent limitations in utilizing long-range dependence and global contextual information, leading to performance degradation in outdoor scenes with many small or blurred objects. To solve this problem, this study proposes a CSTG module that combines a spatial transformer (ST) and a content-aware gating (CAG) mechanism. CSTG refines the feature expression across multiple stages by integrating global context alignment, channel readjustment, and semantic-based selective emphasis into a single lightweight structure. Figure 3 demonstrates the entire configuration flow of CSTG.

 doi:10.20944/preprints202511.2035.v1

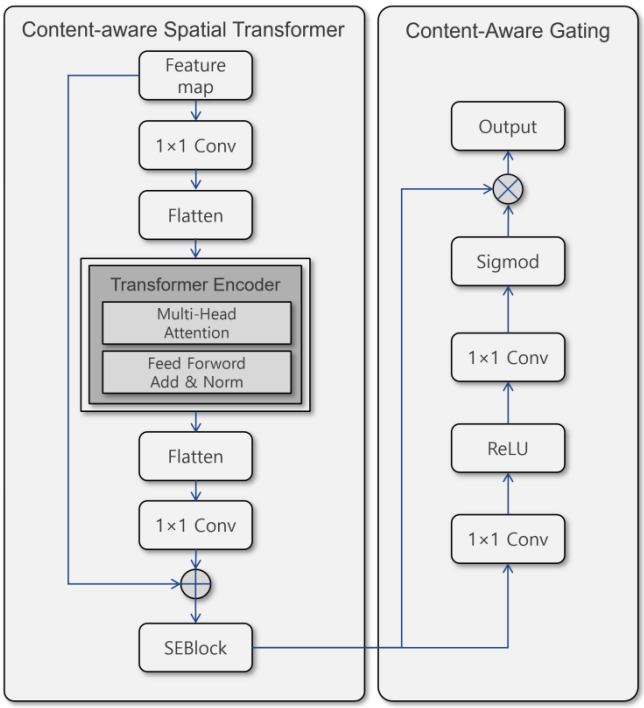

Figure 3. Overall architecture of the CSTG module.

First, the input feature map is reduced in dimension via a 1×1 convolution, then spread across the spatial dimension and entered into the Transformer encoder. The Transformer encoder includes multi-head self-attention, FFN, and residual connections, and models long-distance interactions across the feature map to reconstruct the global context information needed for blurred areas or obscured objects. The Transformer output is restored to its original spatial structure and projected back to the original number of channels via a 1×1 convolution. Subsequently, a Squeeze-and-Excitation Block (SEBlock) is applied to re-importance channels based on the global context [30]. SEBlock summarizes the average response across the entire space and reweights the semantic importance of each channel, thereby suppressing noise channels and enhancing meaningful channels. Figure 4 shows the structure of SEBlock.

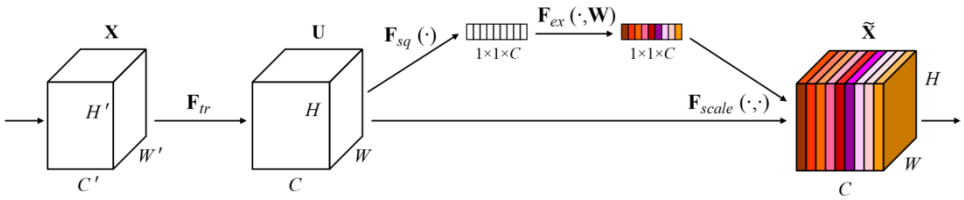

Figure 4. SEBlock used in CSTG for channel re-weighting.

After global refining through Transformer and SE-based channel rearrangement, the content-based semantic selectivity is further secured by the CAG. The CAG consists of two consecutive gating blocks, each composed of a 1×1 convolution followed by ReLU and sigmoid activations. It directly calculates the importance of channel and spatial units based on the regional semantic distribution of input features. Unlike SEBlock, based on the global average, it extracts gating weights directly from the original spatial structure, enabling more detailed emphasis in scenes where small objects or regional semantic changes are important. The operation of CAG is expressed as follows:

$$F' = F \cdot \sigma(Conv_{1\times1}(ReLU(Conv_{1\times1}(F)))) \tag{16}$$

 doi:10.20944/preprints202511.2035.v1

Here, F is a feature map that has undergone Transformer and SEBlock, and F' is the final purification output with the gating module's content-based gate applied.

This structural design enables CSTG to continuously integrate global context, channel importance, and regional semantic information, increase the stability of feature alignment across scales, and make semantic separation between objects and backgrounds more straightforward. Sensitivity is greatly improved, especially in real-world environments with many small, blurred, and partially obscured objects. Despite their lightweight structure, they act as key factors in significantly improving robustness and discrimination in complex outdoor scenes.

3.3. Joint Optimization and Training Objective

In the entire framework proposed in this study, the DIP module that corrects the degraded quality of the input image, the LEGA module that reinforces structural clues, and the CSTG module that performs global-to-regional context alignment are all co-optimized within a single end-to-end learning scheme. At this time, all modules are designed to improve detection performance and do not require additional pre-training or independent auxiliary losses. In other words, all changes in output occurring in each module are backpropagated through the final detection loss, and the parameters of the entire network are jointly updated.

First, the DIP module applies differentiable filters using the parameters predicted by the mini-CNN, and the resulting enhanced image is then transmitted to the subsequent SSD Backbone and Extra Layers. Since the DIP parameter is fine-tuned to enhance object detection performance rather than a fixed rule-based transformation, the entire image quality improvement process is optimized as a task-driven process. The LEGA module highlights structural boundary cues in the backbone's show stage, improving the ability to detect small objects and blurry boundaries, and CSTG performs global context-based refinement and content-based gating in Extra Layers to enhance feature alignment and semantic separation across scales.

The target function of the entire network follows the SSD-based standard detection loss. In the training process, classification loss $\mathcal{L}_{cls}$ and bounding box regression loss $\mathcal{L}_{loc}$ are calculated at the same time, and the final objective function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{loc} \tag{17}$$

Since the DIP, LEGA, and CSTG modules proposed in this study are all completely differentiable and directly connected to detection loss, all parameters are updated together by a single objective function as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L} \tag{18}$$

Here, θ is a set of parameters of the entire network, including all of the DIP parameters, convolution weights, transformer-based parameters, and gating parameters.

Through this integrated training procedure, the three steps of degradation correction-structural emphasis-contextual alignment work complementarily, providing much higher consistency and reliability than the way individual modules are designed independently. As a result, the proposed model can achieve robust detection performance even under adverse weather conditions and maintain stable detection performance across objects of varying scales and complex scene structures.

4. Experiments

4.1. Experimental Setup

4.1.1. Hardware and Software Environment

All experiments were conducted in the Ubuntu 20.04.6 LTS environment, and the server is equipped with 8 NVIDIA Tesla V100 GPUs with 32GB of memory and AMD EPY 7742 processors. The

 doi:10.20944/preprints202511.2035.v1

training code was implemented in Python 3.12, and the model was trained using PyTorch 2.2.1 and CUDA 12.2. Table 5 summarizes the main components of the experimental environment.

Table 5. Experimental Environment Configuration.

Category	Specification
Operating System	Ubuntu 20.04.6 LTS (Focal Fossa)
Processor (CPU)	AMD EPYC 7742 64-Core Processor × 2
Memory (RAM)	1.0 TiB DDR4 ECC
GPU Model	NVIDIA Tesla V100-PCIE-32GB × 8
CUDA Version	12.2
PyTorch Version	2.2.1 (GPU-accelerated)
Language	Python 3.12

4.1.2. Dataset Preparation

The nuScenes dataset was used to evaluate the robustness under reverse-weather conditions [31]. nuScenes contains real self-driving data collected in Singapore and Boston and includes various weather conditions (rain, fog), illumination conditions (day, night), and complex urban traffic conditions. It is also a large multimodal dataset that provides 6 cameras, a 360-degree LiDAR, and 5 radars.

In this study, two types of object detection, pedestrian and vehicle, were targeted, and a Filter-nuS subset was constructed by selecting only images corresponding to adverse weather conditions. The entire configuration is shown in Table 6.

Table 6. Filtered dataset derived from nuScenes.

Dataset	Split	Number of Images	Conditions	Classes
Filter-nuS	Train Test	9,950 900	Rain, Fog, Low Light	Pedestrian, Vehicle

4.1.3. Training Strategy

The Stochastic Gradient Descent (SGD) optimizer is used for model training, with an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 5×10^{-4}. A linear warm-up is applied during the first 500 iterations, after which the learning rate follows the step-wise decision schedule. The entire learning was done with 200,000 iterations.

Training was conducted with a batch size of 32 in an 8-GPU environment, and the unbalanced anchor distribution was mitigated via hard negative mining (N:P = 3:1), accounting for the characteristics of SSD-based detectors. The prior box follows the basic SSD configuration and uses a predefined aspect ratio across six feature levels.

The loss function consists of Smooth L1 regression and Softmax classification losses, as in the existing SSD. Losses are calculated for positive samples and selected hard negatives, and the total loss change is shown in Figure 5.

 doi:10.20944/preprints202511.2035.v1

Figure 5. Training loss curve of the proposed DLC-SSD model on the Filter-nuS dataset.

In the inference stage, Non-Maximum Suppression (NMS) with an IoU threshold of 0.45 is applied, and the confidence threshold is set to 0.01 to preserve low-confidence objects that may occur in bad-weather environments.

4.1.4. Evaluation Metrics

The performance of the proposed model is evaluated using Mean Average Precision (mAP), a standard in the object detection field. Precision and Recall definitions are as follows:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN} \tag{19}$$

For each class, AP is computed as the area under the Precision-Recall curve, and the total mAP is defined as the average across all classes.

$$AP = \int_0^1 P(R)dR \approx \sum_n (R_n - R_{n-1})P_n \tag{20}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{21}$$

Here, P and R denote precision and recall, respectively. N is the number of classes, and AP_i is the average precision for class i. The IoU calculation formula is as follows:

$$IoU = \frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})} \tag{22}$$

This indicator is suitable for quantitatively evaluating the performance of small objects in bad weather environments and boundary preservation in low-light environments.

4.2. Experimental Results and Analysis

4.2.1. Quantitative Comparison with Baselines

This section evaluates the quantitative performance under bad weather conditions by comparing the proposed model with several recent object detection baselines. For comparison, the mean mAP (Mean-mAP), pedestrian mAP (P-mAP), and vehicle mAP (V-mAP) were measured on the Filter-nuS test set, and the results are summarized in Table 7.

 doi:10.20944/preprints202511.2035.v1

Table 7. Comparison of detection performance (mAP) of different methods on the Filter-nuS dataset.

Group	Method	Mean-mAP	P-mAP	V-mAP	Loss	Params
Baseline	D-Yolo	62.91%	54.9%	70.8%	0.51	46M
	ClearSight-OD	61.81%	49.2%	71.5%	0.54	60M
	Aug-Yolov5	62.53%	50.6%	74.4%	0.53	21M
	AK-Net	63.70%	52.1%	73.0%	0.50	75M
	ZeroDCE	59.67%	48.1%	70.5%	0.51	25M
	DSNet	62.10%	48.2%	75.8%	0.52	40M
Ours	CSTG-SSD	62.54%	50.4%	74.6%	0.49	28M
	DLC-SSD	64.29%	53.5%	75.0%	0.44	29M

As shown in Table 7, the proposed model achieved the best performance among all comparison methods, with a Mean-mAP of 64.29%. In particular, it attained high accuracy in pedestrian detection at 53.5% and in vehicle detection at 75%, both of which are susceptible to blurring under adverse weather conditions.

The existing image enhancement-based technique, ZeroDCE, effectively improves input image quality, but its improvement in detection performance is limited due to a lack of specialized optimization for object detection. In addition, models dedicated to bad-weather detection, such as D-YOLO, DSNet, ClearSight-OD, and AK-Net, showed lower mean-mAP than the proposed method, despite their complex structures based on feature adaptation or restoration networks.

On the other hand, the proposed model obtains multi-filter-based differentiable improvement in response to deterioration factors such as brightness, color, tone, and noise through the DIP module, strengthens the structural boundary of shallow stages through LEGA, and improves global context and semantic selectivity through CSTG in an integrated manner within the SSD-based lightweight structure. In particular, it is confirmed that the total number of parameters is 29M, which is relatively lightweight compared to existing restoration-combination models and exhibits high parameter efficiency relative to performance.

4.2.2. Qualitative Results

To further assess the effectiveness of the proposed framework under adverse weather conditions, we conducted a qualitative comparison on the Filter-nuS test split. Figure 6 includes an example of visually comparing the results generated by several detection models under various deterioration conditions, such as rain, fog, and low illumination, on the Filter-nuS test set.

Figure 6. Qualitative visualization of detection results on the Filter-nuS test set.

 doi:10.20944/preprints202511.2035.v1

The existing method frequently blends pedestrians and vehicles with the background because the object outlines cannot be clearly distinguished when the boundaries are blurred. In particular, in scenes where fog or non-structures overlap, a phenomenon in which small objects are completely omitted due to the decrease in contrast of the original image and structural loss, or, on the contrary, a change in brightness of the background is incorrectly detected. In low-light scenes, semantic features were not sufficiently expressed due to insufficient illumination, making it challenging to locate and classify objects accurately.

In contrast, the proposed model helps re-expose the basic structure of an object that appeared blurry by stably correcting brightness, color, contrast, and noise via the DIP module in the input stage. LEGA, which is then applied at the show stage, uses Laplacian-based structural signals to strengthen the outlines and boundaries of objects, more clearly reconstructing the silhouettes of small objects or those that have collapsed in shape. In addition, CSTG stabilizes overall scene representation by using global contextual information to more clearly emphasize semantically important areas and to suppress background deterioration factors relatively.

This series of processing results also shows a clear visual difference. In the proposed model, the boundary between the object and the background is clear even in cloudy scenes, and the shapes of difficult-to-recognize objects, such as small pedestrians or distant vehicles, become more evident. The example in Figure 6 shows that the proposed framework simultaneously maintains structural, color, and semantic balance under adverse weather conditions and can stably detect objects even under various deterioration conditions.

4.2.3. Ablation Study

An ablation experiment was performed to quantitatively confirm the role of each component of the proposed DLC-SSD framework in the final detection performance. The experiment used the Filter-nuS test set, and performance changes were measured by sequentially combining the DIP, LEGA, and CSTG modules with the SSD basic structure. The overall results are summarized in Table 8.

Table 8. Ablation study of the proposed components on the Filter-nuS dataset.

Enhancement	**Detector**				**Performance**	
DIP	SSD	LEGA	ST	CAG	mAP	Times
	✓				61.01%	12ms
	✓		✓	✓	61.95%	14ms
	✓	✓	✓	✓	62.54%	15ms
✓	✓				63.65%	15ms
✓	✓	✓	✓	✓	64.29%	19ms
		Ours vs. SSD			3.28%↑	-

A pure SSD model with no enhancement module applied records a mAP of 61.01% as a reference point for basic performance. Applying the CSTG structure, including ST and CAG, slightly increases performance to 61.95%, demonstrating that Transformer-based global contextual information and gating-based semantic filtering improve the expressive power of the deep layer. Although it does not include image restoration or structure enhancement, a performance improvement of about 1%p was observed, with only semantic-level improvement. Adding LEGA to CSTG resulted in higher performance at 62.54%. Since LEGA reinforces boundary and outline information by extracting Laplacian-based structural information from the show feature, the expressive power of objects with blurred boundaries, especially in bad weather conditions, is improved.

On the other hand, the DIP-SSD structure that uses only the DIP module achieves an mAP of 63.65%. By directly suppressing noise and correcting contrast and tone at the image stage, potential structural information of the object is reconstructed more clearly at the input stage, resulting in a

 doi:10.20944/preprints202511.2035.v1

sufficiently significant performance improvement even without feature-level improvement. However, due to the lack of semantic-level adjustment (CSTG) or shallow-edge correction (LEGA), a limit in which feature alignment is not sufficiently stable in certain scenes is also observed.

Finally, the overall model (DLC-SSD), which integrates DIP, LEGA, and CSTG, achieved the highest performance with an mAP of 64.29%. This is 3.28%p higher than the basic SSD, 1.75%p higher than the CSTG-SSD, and 0.64%p higher than the DIP-SSD, and clearly shows that the multi-layer correction in the image-structure-meaning stage works in a complementary manner. In particular, the performance difference was more pronounced in subcategories with a high proportion of small objects, meaning that the combination of LEGA and DIP enables stable contour restoration even in structurally weak inputs.

Overall, ablation experiments demonstrate that each module provides meaningful improvements on its own, but the highest consistency and robustness are achieved when the three modules are integrated. This is the key basis for showing that the proposed multi-level unified refinement strategy is not a simple module combination but a practical synergy effect enabled by a complementary structure.

4.2.4. Efficiency Analysis

This section quantitatively verifies the lightweight and real-time processability of the proposed DLC-SSD framework as can be seen from Table 7, the total number of parameters of the proposed model is about 29M, which is significantly less than conventional models for adverse weather response, such as D-YOLO (46M), ClearSight-OD (60M), and AK-Net (75M). This is the result that the DIP, LEGA, and CSTG modules are all designed to improve performance without significantly changing the basic structure of the SSD based on a lightweight design.

The DIP module includes various differentiable filters, but each filter is configured as a single operation, so there is little increase in parameters, and the number of operations is also limited because a small mini-CNN predicts the filter parameters. LEGA also maintains a structure with a few additional parameters, using only fixed Laplacian operations and 1×1 convolutions, and CSTG is designed with a 1-layer Transformer encoder and a simple gating block, which is very efficient compared to existing attitude-based models.

The inference speed also supports the lightweight nature. In the ablation study of Table 8, the reference time of the proposed model was measured as 19ms, which is only a slight increase compared to the existing SSD. This efficiency is possible because the DIP operates only in the image input stage, and the CSTG and LEGA are designed to be applied only to a specific layer, not to the entire multilayer feature.

Taken together, the proposed DLC-SSD model maintains the lightweight, high-efficiency structure of the SSD-based model despite including additional modules, and provides performance and processing speed that can be applied immediately, even in a harsh weather monitoring environment where real-time object detection is required. This confirms its value as a practical framework that secures both accuracy and efficiency.

5. Conclusions

To address the significant deterioration in object detection performance in bad-weather environments, this study proposed a DLC-SSD framework that achieves lightweight, real-time performance. The proposed model is fundamentally different from the existing approach in that it organically connects the three layers of image-structure-semantic and directly optimizes the entire process from input image correction to feature alignment within a single end-to-end structure.

First, DIP, the Differentiable Image Processing module, configures all image-correction filters in a fully differentiable form and operates in a way that enables a mini-CNN to predict filter parameters. Through this, task-driven enhancement was realized, directly optimized by detection loss without the need for a separate restoration network or a correct answer image. This method effectively restores

 doi:10.20944/preprints202511.2035.v1

potential signals based on brightness, color, tone, and weather conditions, providing more stable input expressions in low-quality images.

Second, LEGA, the Lightweight Edge-Guided Attention module, uses a lightweight depthwise convolution with a fixed Laplacian kernel to reinforce boundary and outline information that is prone to loss in the high-resolution shallow feature map of SSD. LEGA, which operates without additional parameters, mitigates structural degradation peculiar to bad weather, such as blurred boundaries, small objects, and partial coverings, to ensure structural stability in the early stages.

Third, CSTG, the Content-aware Spatial Transformer with Gating, combines Transformer-based global spatial context processing and CAG to effectively reduce local-global information imbalances caused by irregular lighting, precipitation, scattering, etc. This strengthens semantic matching across multiple scales, suppresses unnecessary background noise, and selectively emphasizes object-related channels to maintain robust expression even in adverse weather conditions.

Each of these three modules independently demonstrates performance improvement, but the synergy is most significant when integrated into a single improvement flow that encompasses all image, structure, and semantic layers. In the Filter-nuScenes dataset-based experiment, the proposed model achieved an average mAP of 64.29%, showing a 3.28%p improvement over the existing SSD and outperforming all D-YOLO, DSNet, and ZeroDCE methods. The practical value of the approach for actual monitoring, transportation, and autonomous driving systems was also confirmed, as it maintains high performance even under various adverse weather conditions.

Nevertheless, some limitations remain. First, since the DIP module is indirectly optimized by detection loss, changes in scene context or direct adaptation to specific weather patterns may still be limited. Second, since LEGA is an optimized structure for shallow features, the effect of restoring structural information at a high level is relatively weak. Third, since this study is designed around a single-frame condition based on 2D images, additional studies are required in scenes where temporal continuity or 3D context is important.

Future research may investigate weather-aware adaptive enhancement, in which an explicit estimation of scene-specific weather conditions dynamically guides the DIP parameters. Another promising direction is multimodal fusion, incorporating additional sensing modalities such as LiDAR, radar, or thermal imaging, to further improve robustness under severe visibility degradation. Extending the framework toward temporal modeling could enhance detection stability across consecutive frames in video sequences. Finally, exploring model compression and neural architecture search techniques would enable more efficient deployment on resource-constrained edge devices without compromising detection performance.

Overall, DLC-SSD is an integrated detection framework that is resistant to video deterioration caused by bad weather while maintaining real-time performance and lightness, and has high potential for use in various real-world vision systems, such as road monitoring, intelligent transportation, robots, and autonomous driving.

Author Contributions: Conceptualization, J.K., C.J. and Y.S.; methodology, S.P., J.K., C.J. and Y.S.; software, J.K. and H.Kim; validation, S.P., J.K. and Y.S.; formal analysis, J.K. and C.J.; investigation, S.P., H.Kim and H.Ko; resources, J.K. and H.Ko; data curation, S.P. and J.K.; writing—original draft preparation, S.P., J.K. and H.Kim; writing—review and editing, S.P., C.J. and Y.S.; visualization, S.P., J.K. and H.Kim; supervision, Y.S.; project administration, C.J. and Y.S.; funding acquisition, C.J. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT) (2710086167). This work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT) (RS-2025-02412990).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202511.2035.v1

Data Availability Statement: The nuScenes dataset is available at https://www.nuscenes.org/nuscenes. (accessed on 23 November 2025) [31].

Acknowledgments: This paper was supported by Wonkwang University in 2025.

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

References

1. Seo, A.; Woo, S.; Son, Y. Enhanced Vision-Based Taillight Signal Recognition for Analyzing Forward Vehicle Behavior. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24165162.
2. McCartney, E.J.; Hall, Freeman F., J. Optics of the Atmosphere: Scattering by Molecules and Particles. *Physics Today* **1977**, *30*, 76–77, [https://pubs.aip.org/physicstoday/article-pdf/30/5/76/8283978/76_1_online.pdf]. https://doi.org/10.1063/1.3037551.
3. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1956–1963. https://doi.org/10.1109/CVPR.2009.5206515.
4. Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; Yang, M.H. Multi-Scale Boosted Dehazing Network With Dense Feature Fusion. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2154–2164. https://doi.org/10.1109/CVPR42600.2020.00223.
5. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7313–7322. https://doi.org/10.1109/ICCV.2019.00741.
6. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1777–1786. https://doi.org/10.1109/CVPR42600.2020.00185.
7. Kim, Y.T.; Bak, S.H.; Han, S.S.; Son, Y.; Park, J. Non-contrast CT-based pulmonary embolism detection using GAN-generated synthetic contrast enhancement: Development and validation of an AI framework. *Computers in Biology and Medicine* **2025**, *198*, 111109. https://doi.org/https://doi.org/10.1016/j.compbiomed.2025.111109.
8. Kim, H.; Son, Y. Generating Multi-View Action Data from a Monocular Camera Video by Fusing Human Mesh Recovery and 3D Scene Reconstruction. *Applied Sciences* **2025**, *15*. https://doi.org/10.3390/app151910372.
9. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
11. Huang, S.C.; Le, T.H.; Jaw, D.W. DSNet: Joint Semantic Learning for Object Detection in Inclement Weather Conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *43*, 2623–2633. https://doi.org/10.1109/TPAMI.2020.2977911.
12. Chu, Z. D-YOLO a robust framework for object detection in adverse weather conditions. *arXiv preprint arXiv:2403.09233* **2024**. https://doi.org/10.48550/arXiv.2403.09233.
13. Le, T.H.; Huang, S.C.; Hoang, Q.V.; Lokaj, Z.; Lu, Z. Amalgamating Knowledge for Object Detection in Rainy Weather Conditions. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*. https://doi.org/10.1145/3712703.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision – ECCV 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds., Cham, 2016; pp. 21–37.
15. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems* **2019**, *30*, 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865.
16. Liu, B.; Jin, J.; Zhang, Y.; Sun, C. WRRT-DETR: Weather-Robust RT-DETR for Drone-View Object Detection in Adverse Weather. *Drones* **2025**, *9*. https://doi.org/10.3390/drones9050369.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.

 doi:10.20944/preprints202511.2035.v1

18. Wang, Y.; Zhang, J.; Zhou, J.; Han, M.; Li, S.; Miao, H. ClearSight: Deep Learning-Based Image Dehazing for Enhanced UAV Road Patrol. In Proceedings of the 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2024, pp. 68–74. https://doi.org/10.1109/CVIDL62147.2024.10603766.
19. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions **2022**. [arXiv:cs.CV/2112.08088].
20. Kalwar, S.; Patel, D.; Aanegola, A.; Konda, K.R.; Garg, S.; Krishna, K.M. GDIP: Gated Differentiable Image Processing for Object-Detection in Adverse Conditions **2022**. [arXiv:cs.CV/2209.14922].
21. Ogino, Y.; Shoji, Y.; Toizumi, T.; Ito, A. ERUP-YOLO: Enhancing Object Detection Robustness for Adverse Weather Condition by Unified Image-Adaptive Processing **2024**. [arXiv:cs.CV/2411.02799].
22. Gupta, H.; Kotlyar, O.; Andreasson, H.; Lilienthal, A.J. Robust Object Detection in Challenging Weather Conditions. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7508–7517. https://doi.org/10.1109/WACV57701.2024.00735.
23. Ghiasi, G.; Fowlkes, C.C. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation **2016**. [arXiv:cs.CV/1605.02264].
24. Cai, J.; Sun, H.; Liu, N. B2Net: Camouflaged Object Detection via Boundary Aware and Boundary Fusion **2024**. [arXiv:cs.CV/2501.00426].
25. Bui, N.T.; Hoang, D.H.; Nguyen, Q.T.; Tran, M.T.; Le, N. MEGANet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation **2023**. [arXiv:cs.CV/2309.03329].
26. Qiu, H.; Ma, Y.; Li, Z.; Liu, S.; Sun, J. BorderDet: Border Feature for Dense Object Detection **2021**. [arXiv:cs.CV/2007.11056].
27. Gharatappeh, S.; Sekeh, S.; Dhiman, V. Weather-Aware Object Detection Transformer for Domain Adaptation **2025**. [arXiv:cs.CV/2504.10877].
28. Jiang, L.; Ma, G.; Guo, W.; Sun, Y. YOLO-DH: Robust Object Detection for Autonomous Vehicles in Adverse Weather. *Electronics* **2025**, *14*. https://doi.org/10.3390/electronics14224476.
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition **2015**. [arXiv:cs.CV/1409.1556].
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
31. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.