CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Task-Aligned Haze Removal With Semantic-Aware Fusion and Contrast Self-Correction

Jinbin Wang[1] | Aiping Yang[2] | Guosong Jiang[1] | Wenlong Yu[3] | Dongwei Ren[1] | Qinghua Hu[1]

[1]School of Artificial Intelligence, Tianjin University, Tianjin, China | [2]School of Electrical and Information Engineering, Tianjin University, Tianjin, China | [3]School of Computing and Information Systems, Singapore Management University, Singapore

Correspondence: Aiping Yang (yangaiping@tju.edu.cn)

Received: 15 December 2025 | **Revised:** 25 May 2026 | **Accepted:** 4 June 2026

Keywords: contrast self-correction | haze removal | low-level vision | semantic fusion | task-aligned

ABSTRACT

Adverse haze conditions introduce complex degradations that obscure scene details and distort structural cues critical for object detection, posing persistent challenges for vision-based sensing systems. Although existing haze removal methods have achieved notable improvements in visual clarity, their optimisation objectives are often misaligned with downstream detection requirements, leading to limited detection performance in real-world scenarios. To address this issue, this work proposes a task-aligned weakly supervised haze removal framework, termed Dehaze4Detection, which explicitly aligns low-level restoration with high-level detection objectives. The framework incorporates a Semantic-Aware Multi-Scale Fusion Module (SMFM) that embeds pixel-level semantic knowledge into the dehazing process, enabling selective enhancement of object regions while suppressing background over-enhancement to achieve detection-friendly restoration. To further improve feature quality, a Contrast Self-Correction Attention Module (CSAM) is introduced, which exploits contrast maps derived from pseudo-dehazed images to guide attention towards object contours and salient structures. The proposed SMFM and CSAM are integrated into a unified pipeline and jointly optimised using detection and dehazing losses. Extensive experiments on both synthetic and real-world benchmarks demonstrate that Dehaze4Detection achieves competitive perceptual quality while significantly improving detection accuracy.

1 | Introduction

Haze is a common atmospheric phenomenon caused by the scattering of light from suspended particles in the air. It reduces scene radiance and introduces airlight, resulting in low contrast, distorted colours and loss of structural details. These degradations not only deteriorate visual perception but also interfere with the performance of high-level vision tasks, such as object detection [1–5], as illustrated in Figure 1. The widely used atmospheric scattering model typically describes the physical formation of haze:

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

where $I(x)$ is the observed hazy image, $J(x)$ is the latent clean image, $t(x)$ denotes the transmission map indicating the portion of light that reaches the camera and A is the global atmospheric light. Image dehazing aims to invert this process, recovering the latent clean image $J(x)$ from the observed hazy image $I(x)$. Image dehazing has long been regarded as a critical problem in the field of low-level vision.

Classical single-image dehazing methods aim to recover clear images with enhanced perceptual quality: Early approaches relied on handcrafted priors to inversely solve the atmospheric scattering model with prior constraints, such as the dark

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.



FIGURE 1 | Effect of haze on object detection. The figure illustrates that perception-oriented dehazing methods primarily enhance visual quality but hardly translate into better detection performance. The proposed Dehaze4Detection framework aligns dehazing with downstream detection tasks. (a) Hazy, (b) DNDM [6], (c) SDA-GAN [7], (d) Ours.

channel prior (DCP) [8], colour attenuation prior (CAP) [9] and haze-lines prior [10]. Subsequently, CNN-based models [11–14] integrated physical knowledge into learnable frameworks, such as DehazeNet [11] and AOD-Net [12]. Recently, Transformer-based methods further improved perceptual quality, benefiting from their strong ability to capture long-range dependencies [15–19]. Although the aforementioned perception-oriented methods have achieved impressive visual quality, their improvements hardly translate into better detection performance. This gap stems from the fact that perceptual metrics (e.g., PSNR, SSIM) focus on global pixel-wise similarity rather than the fidelity of task-relevant structures.

More recently, task-aware dehazing methods [20, 21] have been introduced to improve detection performance under haze by jointly optimising dehazing and detection objectives typically through shared backbone features or task-driven loss functions. These methods have achieved considerable performance gains and demonstrate the effectiveness of task-driven optimisation. However, most of them overlook the value of pixel-level semantic knowledge. Without such semantic guidance, these models tend to under-enhance critical object regions while over-enhancing irrelevant background areas, leading to structural distortion and degraded detection performance. Pixel-level semantic knowledge could provide spatial and category cues directly aligned with detection objectives, enabling the model to distinguish foreground objects from background regions and perform targeted enhancement. For instance, in scenes where haze density is nonuniform, such semantics knowledge allows the model to prioritise restoring semantically important regions (e.g., pedestrians, vehicles) while avoiding unnecessary enhancement of irrelevant backgrounds.

To this end, we propose a task-aligned weakly supervised dehazing framework, termed Dehaze4Detection, which explicitly aligns image restoration with downstream detection tasks. The framework is highlighted by a Semantic-Aware Multi-Scale Fusion Module (SMFM) that integrates pixel-level semantic knowledge, which can be obtained from a pretrained segmentation model into the dehazing reconstruction process. By adaptively fusing these semantic cues across multiple scales, SMFM enhances structural fidelity and suppresses artefacts, ensuring that the restored features are semantically consistent and detection-friendly. Although SMFM provides pixel-level semantic enhancement, effective dehazing reconstruction also depends on the quality of the extracted features used as input. Under adverse conditions where contrast is severely degraded and object contours appear blurred, the extracted features may become corrupted or lack sufficient discrimination, thus hindering the recovery of clean and consistent structures. In such cases, direct contrast enhancement often amplifies noise or introduces artefacts. To overcome these limitations, we introduce a Contrast Self-Correction Attention Module (CSAM) that refines the model's attention towards high-contrast regions typically aligned with object contours and salient structures. This correction suppresses artefacts and enhances the sharpness of object contours, thereby improving feature quality and strengthening detection robustness. Finally, the proposed SMFM and CSAM are integrated into a unified pipeline and optimised under a joint detection and dehazing loss. The main contributions of this work are summarised as follows:

- This work introduces Dehaze4Detection, a task-aligned weakly supervised image dehazing framework that jointly optimises image restoration and detection objectives. Unlike perception-driven approaches that focus solely on visual quality, Dehaze4Detection is carefully designed to bridge the gap between low-level enhancement and high-level detection tasks by leveraging task-relevant information throughout the reconstruction process. This alignment ensures that restored images are not only visually pleasing but also structurally consistent and detection-friendly.
- Two novel modules are proposed to effectively realise the framework. The first, a Semantic-Aware Multi-Scale Fusion Module (SMFM), is highlighted for embedding pixel-level semantic knowledge extracted from a pretrained segmentation model into the reconstruction pipeline, enabling the network to selectively enhance object regions and suppress over-enhancement in background areas. The second, a Contrast Self-Correction Attention Module (CSAM), cleverly leverages contrast maps from pseudo-dehazed images to guide attention towards object contours and salient regions, thereby reducing artefacts and improving structural clarity.
- Extensive experiments are conducted on both synthetic and real-world benchmarks, including RESIDE (SOTS, HSTS, RTTS) [22] and Foggy Cityscapes [23]. Experimental results show that Dehaze4Detection achieves competitive perceptual quality (in PSNR/SSIM) while significantly outperforming prior methods in detection accuracy. In particular, Dehaze4Detection increases mAP@0.5 on RTTS from 61.6% to 63.7% and on Foggy Cityscapes from 56.1% to 57.3%. Ablation studies further verify the effectiveness of both SMFM and CSAM as well as the joint optimisation strategy with detection loss.

The remainder of this article is organised as follows. Section 2 presents a review of related work on image dehazing. Section 3 elaborates the proposed Dehaze4Detection framework, including the overall architecture and two core modules. Section 4 reports extensive experiments, ablation studies and comparative analysis

with state-of-the-art approaches. Finally, Section 5 concludes this article and discusses possible directions for future research.

2 | Related Work

This section reviews three lines of research relevant to this work: deep learning–based dehazing, weakly supervised and unsupervised dehazing and task-aware dehazing for high-level vision.

2.1 | Deep Learning–Based Dehazing

Early learning–based dehazing frameworks often coupled atmospheric scattering models with CNN-based estimators. Representative examples include DehazeNet [11], which regresses transmission maps and applies model inversion to restore haze-free images, showing good performance on synthetic datasets but limited generalisation to real-world distributions with dense or uneven haze layers. MSCNN [24] further adopts a coarse-to-fine strategy to refine transmission estimation across scales, improving edge consistency but remaining susceptible to error propagation over homogeneous regions. AOD-Net [12] re-parameterises the physical model by embedding both transmission and airlight into a unified formulation, enabling compact architectures suitable for integration into downstream pipelines though its expressiveness remains constrained under complex illumination conditions. Later works discard explicit atmospheric modelling in favour of end-to-end learning, leveraging richer contextual information to enhance visual realism. For instance, EPDN [25] introduces GAN-based refinement to synthesise high-frequency textures but may induce artefacts around strong gradients. GCANet [26] incorporates smoothed dilated convolutions with gated fusion to better capture global structure, whereas GridDehazeNet [27] and MSBDN-DFF [28] integrate multi-scale attention and dense fusion modules to progressively restore scene details. Despite these improvements, most approaches remain heavily reliant on synthetic datasets and visual metrics, thus facing notable domain shift when deployed in real-world settings. More recent studies introduce attention mechanisms and transformer architectures to address spatial variability in haze. FFA-Net [29] utilises joint pixel and channel-level attention to emphasise informative regions, whereas AECR-Net [30] employs feature-level contrastive constraints to better separate haze-free features from degraded features. Recently, Transformer-based methods further improved perceptual quality, benefiting from their strong ability to capture long-range dependencies [15, 31]. For instance, DehazeFormer [15] achieves state-of-the-art restoration performance via lightweight normalisation blocks and hierarchical transformer architecture, which sets a new benchmark for perception-oriented dehazing. In parallel, dehazing-specific transformers such as Dehamer [31] report additional gains through transmission-aware position encoding. Nonetheless, the aforementioned perception-oriented methods, including Transformer-based approaches, focus on optimising global pixel-wise similarity metrics (e.g., PSNR, SSIM) rather than the fidelity of task-relevant structures. Thus, their impressive visual quality improvements hardly translate into better detection performance, which is the core gap this work aims to bridge.

2.2 | Weakly/Unsupervised Dehazing

Researchers have explored weakly supervised and unsupervised learning paradigms to mitigate the scarcity of real-world paired data. RefineDNet [32] divides dehazing into two stages: visibility recovery guided by handcrafted priors and adversarial refinement with unpaired real images. Although it enhances realism, artefacts from prior initialisation may persist. DMT [33] introduces a mean-teacher framework with disentangled representations (e.g., radiance, transmission and airlight), leveraging consistency constraints across domains, though the dual-branch setup increases training instability. Unsupervised methods further eliminate the need for annotations. DCP-loss [34] training imposes handcrafted priors as regularisation terms in the learning objective, whereas ZID [35] decomposes a hazy image into constituent physical components via dedicated subnetworks, allowing per-image adaptation at inference time. Cycle-consistent adversarial frameworks, such as Cycle-Dehaze [36], apply unpaired translation between clear and hazy domains but may introduce tonal shifts or blurred textures depending on content complexity. Prior-free approaches such as YOLY [37] rely on internal statistics to guide restoration without any external supervision, though performance significantly degrades under dense or heterogeneous haze. Domain adaptation methods tackle the synthetic–real discrepancy via explicit distribution alignment. DA-Dehaze [38] integrates bi-directional image translation with prior-guided modules to align feature spaces, whereas source-free UDA [39] adapts pre-trained source models to unseen target domains via normalisation and frequency-domain constraints without revisiting source data. These strategies have reduced reliance on paired samples and improved generalisation, but they often lack fine-grained semantic alignment particularly around object contours crucial for detection.

2.3 | Task-Aware Dehazing for High-Level Vision

A growing body of work emphasises the importance of coupling image restoration with high-level visual understanding. Rather than optimising solely for PSNR or SSIM, these approaches integrate task-specific objectives, particularly object detection, into the dehazing pipeline. Dual-branch models such as TogetherNet [40] jointly train enhancement and detection branches, sharing feature representations and enabling task-guided restoration. Similarly, Detection-Friendly Dehazing [41] introduces a bilinear attention architecture that fuses hazy and dehazed features through an attention module and employs a self-supervised haze robust loss along with an interval iterative training strategy to achieve detection-friendly dehazing without relying on clear-image labels. Such coupling improves mAP on foggy scenes but can complicate optimisation and reduce interpretability. Extensions such as density-aware pipelines further adapt dehazing and detection to different haze levels, improving robustness across diverse scenes but at the expense of increased network complexity [42, 43]. In parallel, differentiable preprocessing and image-adaptive detection integrate learnable

enhancement modules into detectors to co-optimise recognition under adverse visibility [44, 45]. Other frameworks, such as FriendNet [20], design dedicated guidance modules to fuse task-relevant features into restoration layers, thus achieving a closer alignment between low-level enhancement and semantic prediction. Although these designs can improve detector performance, they may also degrade visual restoration quality and introduce artefacts when task-related features dominate. Alternatively, the work in [46] leverages causal intervention on a confounder dictionary to disentangle object features from weather-specific perturbations, seeking to learn domain-invariant representations for detection under multiple adverse conditions. On a different front, detector-side adaptation methods embed haze priors directly into the detector's domain alignment process [38]. This strategy bypasses the need for explicit dehazing, underscoring the mutual impact between restoration and detection pathways. More recently, USCFormer [17] first introduces pixel-level semantic knowledge into dehazing via semantic contrastive learning, which provides important inspiration for our work. However, USCFormer only leverages semantic maps in the post-processing stage to refine the final output without integrating semantic knowledge into the core dehazing reconstruction pipeline, failing to achieve end-to-end semantic-guided enhancement. Moreover, most existing task-aware methods overlook the value of pixel-level semantic knowledge. Without such semantic guidance, these models tend to under-enhance critical object regions while over-enhancing irrelevant background areas, leading to structural distortion and degraded detection performance. Pixel-level semantic knowledge could provide spatial and category cues directly aligned with detection objectives, enabling the model to distinguish foreground objects from background regions and perform targeted enhancement.

3 | The Proposed Method

3.1 | Overall Framework

The overall architecture of the proposed Dehaze4Detection framework is illustrated in Figure 2. The framework consists of two key components: the Semantic-Aware Multi-Scale Fusion Module (SMFM) and the Contrast Self-Correction Attention Module (CSAM). The SMFM serves as the core component of the framework. The SMFM performs semantic-aware multi-scale enhancement through stacked Semantic-Aware Multi-Scale Fusion Blocks (SMFB, detailed in Figure 3) interleaved with residual dense blocks (RDB) [48]. Each SMFB comprises a semantic-guided branch and a multi-scale spatial context branch, whose outputs are fused via channel attention to strengthen local feature representation. This design enables the network to selectively enhance semantically important object regions while suppressing background over-enhancement, thereby preserving semantic consistency and improving structural fidelity. Although the SMFM provides pixel-level semantic enhancement, effective dehazing reconstruction still requires high-quality extracted features. To address this requirement, the CSAM is applied before semantic enhancement to extract attention features through stacked convolutional blocks and multi-scale attention blocks (MAB, detailed in Figure 4) and refines them using a contrast map computed from a pseudo-

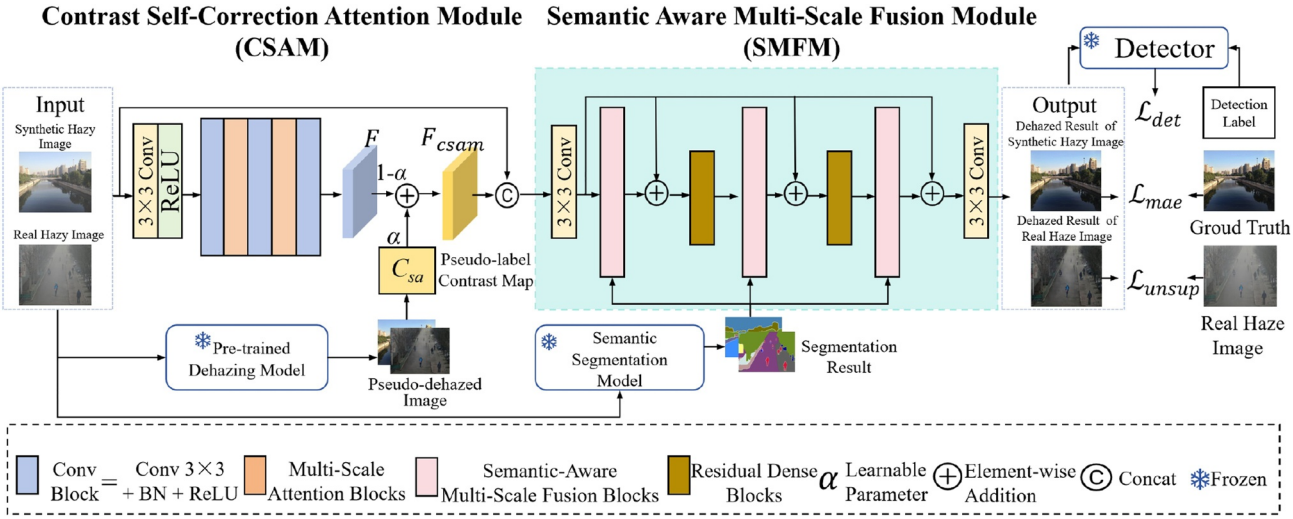


FIGURE 2 | Architecture of the proposed Dehaze4Detection framework. The framework consists of two key components: the semantic-aware multi-scale fusion module (SMFM) and the contrast self-correction attention module (CSAM). The SMFM embeds pixel-level semantic knowledge extracted from a pretrained semantic segmentation model (PSPNet [47]) into the dehazing reconstruction through stacked semantic-aware multi-scale fusion blocks (SMFBs, Figure 3), selectively enhancing object regions while suppressing background over-enhancement. Residual dense blocks (RDBs) [48] are interleaved between SMFBs to strengthen local feature representation through residual–dense connections. The CSAM first extracts attention features with multi-scale attention blocks (MAB, see Figure 4) and then corrects them using contrast maps derived from pseudo-dehazed images produced by a pretrained dehazing model (AOD-Net [12]), thereby enhancing contour sensitivity and suppressing visual artefacts. Convolutional blocks (CB) consist of a 3 × 3 convolution, batch normalisation (BN) and ReLU activation. During training, the dehazed outputs are fed into a fixed detector (YOLOv5s [49]) to compute the detection loss $\mathcal{L}_{\text{det}}$ with detection labels. The optimisation further includes a mean absolute error (MAE) loss $\mathcal{L}_{\text{mae}}$ on synthetic hazy/clean pairs and an unsupervised loss $\mathcal{L}_{\text{unsup}}$ on real hazy images. The integration of SMFM and CSAM within a unified pipeline, jointly optimised with detection and dehazing losses, effectively couples low-level restoration with high-level detection.

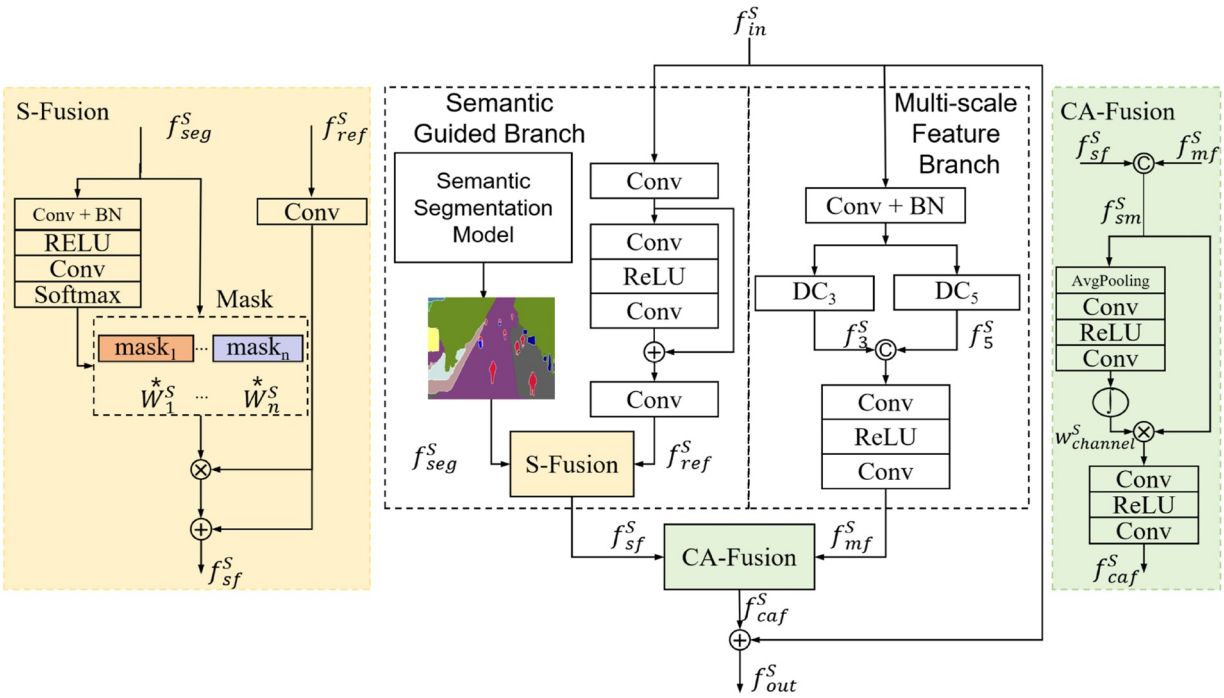


FIGURE 3 | Semantic-aware multi-scale fusion block (SMFB) inside SMFM. SMFB consists of two parallel branches. The semantic-guided branch employs semantic fusion (S-Fusion), which integrates pixel-level semantic knowledge predicted by a pretrained semantic segmentation model (PSPNet [47]) with class-aware weights to selectively emphasise target regions and suppress irrelevant background. The multi-scale spatial context branch applies dilated convolutions with dilation rates 3 and 5 to capture long-range dependencies and preserve local structures. The outputs of the two branches are combined by channel attention fusion (CA-Fusion), which adaptively balances semantic cues with multi-scale spatial features and is finalised with a residual connection. With the combined effect of semantic fusion and channel attention fusion, SMFB improves region-level discriminability and preserves semantic integrity for task-aligned dehazing.

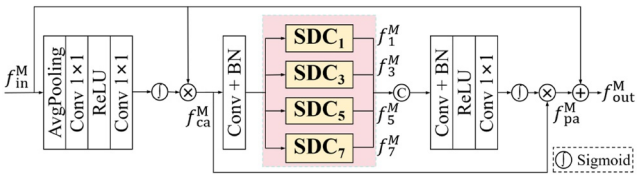


FIGURE 4 | Multi-scale attention block (MAB) inside CSAM. MAB consists of two parallel branches. The channel-attention branch applies global average pooling, 1×1 convolutional layer, ReLU activation and a sigmoid activation to reweight global channel responses. The multi-scale pixel-attention branch applies four smooth dilated 3×3 convolutional layers (SDC) with different dilation rates of $1, 3, 5$ and 7 to extract multi-scale features. The outputs are concatenated and compressed through a 1×1 convolutional layer and a sigmoid gate to generate a pixel-attention map. A residual connection merges the input with the reweighted features, which stabilises gradients and preserves input content. By integrating channel reweighting and multi-scale pixel attention, MAB produces the multi-scale attention features for CSAM.

dehazed image. This correction highlights object contours and salient structures, suppresses artefacts and enhances target–background separability. During training, both synthetic and real hazy images are fed into the network, whereas the dehazed outputs are forwarded to a frozen YOLOv5s [49] detector to compute the detection loss. The entire network is jointly optimised by a hybrid objective that combines detection loss, supervised MAE loss and unsupervised perceptual-quality loss, which are detailed in Section 3.4.

3.2 | Semantic-Aware Multi-Scale Fusion Module (SMFM)

Existing approaches often tend to ignore semantic relationships across different image regions, resulting in incomplete structural recovery in target areas. Moreover, irrelevant regions may be over-enhanced, compromising semantic consistency and hindering the performance of high-level tasks such as object detection. This section addresses these challenges by proposing the Semantic-Aware Multi-Scale Fusion Module (SMFM), which utilises semantic segmentation to dynamically generate class-specific feature weights and corresponding masks, thereby selectively enhancing target regions. In parallel, a multi-scale feature branch captures spatial context, whereas a channel attention mechanism adaptively fuses semantic and contextual features. This design improves the discriminability of target representations and the accuracy of region-level perception. SMFM consists of two key components: Semantic-Aware Multi-Scale Fusion Blocks (SMFB) and residual dense blocks (RDB [48]). Each SMFB comprises two parallel branches, a semantic-guided branch and a multi-scale context branch, whose outputs are fused via a channel attention mechanism. A residual connection is then applied to generate the final output. The

detailed structure is illustrated in Figure 3. In addition, the design employs the standard residual dense block (RDB) [48] without structural modifications to enhance local feature extraction. Each RDB consists of three densely connected convolutional layers and a 1×1 convolutional layer for local fusion.

3.2.1 | *Semantic-Guided Branch*

Let f_{in}^{S} denote the input feature maps to SMFB. In the semantic-guided branch, a stack of convolutional layers is applied to f_{in}^{S}. This design incorporates a short skip connection within the branch, ensuring that local details are preserved whereas deeper-level features are effectively aggregated to obtain a refined feature f_{ref}^{S}:

$$f_{ref}^{S} = \mathrm{C}_3\left(\mathrm{C}_3\left(\delta\left(\mathrm{C}_3\left(\mathrm{C}_3\left(f_{in}^{S}\right)\right)\right)\right) + \mathrm{C}_3\left(f_{in}^{S}\right)\right), \tag{2}$$

here $C_3(\cdot)$ denotes a 3×3 convolutional layer, and $\delta(\cdot)$ is the ReLU activation. The block obtains pixel-level semantic knowledge by applying a pre-trained semantic segmentation model (PSPNet [47]) to the input image I, producing a semantic map f_{seg}^{S} with pixel-wise category predictions:

$$f_{seg}^{S} = segment(I), \tag{3}$$

where $segment(\cdot)$ denotes the semantic segmentation network. The semantic map is decomposed into per-class masks $mask_i$, which are combined with feature f_{ref}^{S} in the Semantic Fusion (S-Fusion). This block further computes class-wise weighting vectors W_i^S to modulate category contributions formulated as:

$$W_i^S = softmax\left(\mathrm{C}_3\left(\delta\left(\mathrm{BN}\left(\mathrm{C}_3\left(f_{seg}^{S}\right)\right)\right)\right)\right), \tag{4}$$

where $softmax(\cdot)$ guarantees a valid probability distribution across classes, $BN(\cdot)$ denotes the Batch Normalisation, $i = 1, \ldots, n$ denotes the number of foreground categories and the background is treated as class 0. The $softmax(\cdot)$ operation in Equation (4) is applied class-wise: for the category prediction score of each pixel position, we perform softmax normalisation along the category dimension, ensuring the sum of weights of all categories (including background class 0) at each pixel position is 1. This implicit normalisation avoids magnitude imbalance between different categories. To explicitly suppress irrelevant background regions, we multiply the weight of the background class (class 0) by a suppression coefficient of 0.1 after the softmax operation, which greatly reduces the contribution of the background and makes the network focus on foreground object regions critical for detection. After the class-wise weighted summation, the subsequent 3×3 convolutional layer and attached Batch Normalisation operation further standardise the feature magnitude, effectively eliminating feature bias caused by overlapping semantic regions or dominant categories in multi-object scenes.

Let $f_{ref}^{S} \in \mathbb{R}^{C \times H \times W}$ and $\mathrm{mask}_i \in \{0, 1\}^{1 \times H \times W}$. Each $W_i \in \mathbb{R}^{C \times 1 \times 1}$ is broadcast to match mask_i and f_{ref}^{S}. The semantic-enhanced feature f_{sf}^{S} is then obtained as:

$$f_{sf}^{S} = \mathrm{C}_3\left(f_{ref}^{S}\right) + \sum_{i=1}^{N} W_i \otimes \mathrm{mask}_i \otimes f_{ref}^{S}, \tag{5}$$

where $\otimes$ denotes element-wise multiplication, $C_3(\cdot)$ denotes a 3×3 convolutional layer. This operation selectively emphasises object-relevant channels and spatial regions while suppressing responses in nontarget areas, thereby improving structural recovery around category-specific contours.

In parallel, a multi-scale feature branch extracts spatial context to complement semantic guidance. Instead of downsampling, two dilated convolutional layers with dilation factors $d = 3$ and $d = 5$ are employed to expand the receptive field without loss of resolution. The dilated feature at each rate d is obtained as:

$$f_d^S = DC_d\left(\mathrm{BN}\left(\mathrm{C}_3\left(f_{in}^{S}\right)\right)\right), \qquad d \in \{3, 5\}, \tag{6}$$

where $DC_d(\cdot)$ denotes a 3×3 dilated convolutional layer with dilation factor d. The resulting features f_3^S and f_5^S are then concatenated and refined to form the multi-scale feature f_m^S:

$$f_m^S = \mathrm{C}_3\left(\delta\left(\mathrm{C}_3\left(\mathrm{concat}\left(f_3^S, f_5^S\right)\right)\right)\right). \tag{7}$$

Finally, a channel-attention fusion (CA-Fusion) block adaptively reweights the semantic and multi-scale features. Taking f_{sf}^S and f_m^S as inputs, it first concatenates them as:

$$f_{sm}^S = \mathrm{concat}\left(f_{sf}^S, f_m^S\right), \tag{8}$$

then generates a channel-wise weight map $w_{channel}^S$:

$$w_{channel}^S = \sigma\left(\mathrm{C}_3\left(\delta\left(\mathrm{C}_3\left(\mathrm{AvgPooling}\left(f_{sm}^S\right)\right)\right)\right)\right). \tag{9}$$

The weighted feature is refined to produce f_{caf}^S:

$$f_{caf}^S = \mathrm{C}_3\left(\delta\left(\mathrm{C}_3\left(w_{channel}^S \otimes f_{sm}^S\right)\right)\right). \tag{10}$$

Adding a residual connection with the input feature f_{in}^S yields the final output of the Semantic-Aware Multi-Scale Fusion Block (SMFB):

$$f_{out}^S = f_{caf}^S + f_{in}^S. \tag{11}$$

Through this design, f_{out}^S integrates both task-aligned semantic guidance and spatial contextual information, thereby improving target feature expressiveness and regional perception.

3.2.2 | *Rationale for RDB Integration*

Unlike simpler basic residual blocks, RDB adopts a residual dense structure where the output of each convolutional layer is directly connected to all subsequent layers within the block. We choose RDB over basic residual blocks for two core reasons tailored to our semantic-aware fusion setting:

1. Dense feature reuse for semantic detail preservation: Haze degradation severely erodes fine-grained structural and semantic boundary information of the image. The dense connection of RDB enables sufficient reuse of multi-scale

local features from shallow layers, which are critical for preserving semantic boundaries and object contours, avoiding the loss of semantic information during feature transmission.

2. Stable semantic consistency under haze degradation: The residual dense structure enhances the expression ability of local features without destroying the injected semantic prior, enabling better fusion of semantically weighted features and spatial context features. Meanwhile, dense feature reuse ensures the network stably maintains semantic consistency under varying haze densities without sudden degradation of feature representation caused by severe haze.

3.3 | Contrast Self-Correction Attention Module (CSAM)

Although SMFM provides pixel-level semantic enhancement, effective dehazing reconstruction also relies on the quality of extracted features. When contrast is degraded and object contours are blurred, the extracted features can be corrupted or lack sufficient discrimination, which impairs the recovery of clean and consistent structures. Although direct contrast enhancement is a straightforward solution, it often amplifies noise and introduces artefacts, degrading both visual quality and detection performance. To address this issue, this section proposes a Contrast Self-Correction Attention Module (CSAM), which leverages contrast maps generated from pseudo-dehazed images to refine the model's attention towards high-contrast regions typically aligned with object contours and salient structures. This correction enhances target–background separability and suppresses visual artefacts, thereby improving feature quality for downstream detection tasks. The detailed structure of CSAM is shown in the Figure 5.

Let I denote the input hazy image. The multi-scale attention branch inside CSAM first extracts the original attention-weighted features F:

$$F = CB\left(MAB(CB(MAB(CB(\delta(C_3(I)))))\right). \tag{12}$$

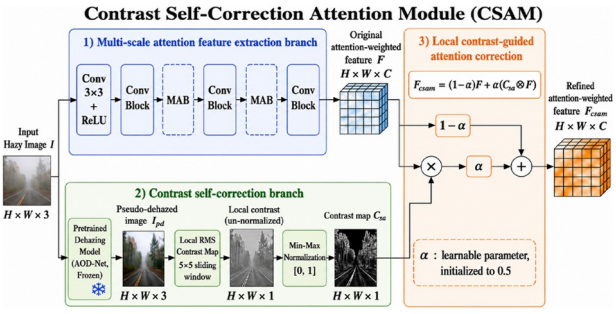


FIGURE 5 | Architecture of the contrast self-correction attention module (CSAM). CSAM consists of a multi-scale attention branch and a contrast self-correction branch. The multi-scale attention branch progressively extracts attention-weighted features through convolutional blocks and multi-scale attention blocks (MABs). CSAM improves feature discriminability and provides more reliable structural representations for subsequent semantic-aware reconstruction.

here $C_k(\cdot)$ denotes a convolutional layer with a kernel size of $k \times k$, $\delta(\cdot)$ is the ReLU activation, $CB(\cdot)$ is the convolutional block defined as a sequential operation consisting of a 3×3 convolutional layer followed by Batch Normalisation and a ReLU activation and $MAB(\cdot)$ denotes the Multi-Scale Attention Block. The Multi-Scale Attention Block (MAB) consists of a channel-attention (CA) branch and a multi-scale pixel-attention (MS-PA) branch. The structure is illustrated in Figure 4. Let f_{in}^M be the input feature. In the first stage, the channel-attention–weighted feature f_{ca}^M is obtained as:

$$f_{ca}^M = \sigma\left(C_1\left(\delta\left(C_1\left(\text{AvgPooling}\left(f_{\text{in}}^M\right)\right)\right)\right)\right) \otimes f_{\text{in}}^M, \tag{13}$$

where AvgPooling$(\cdot)$ denotes the global average pooling operator, $\sigma(\cdot)$ denotes the sigmoid activation, $\delta(\cdot)$ is the ReLU activation, $C_1(\cdot)$ represents a 1×1 convolutional layer and $\otimes$ indicates element-wise multiplication. In the second stage, multi-scale pixel attention is computed. Four smooth dilated convolutional layers with different dilation factors are first applied to produce multi-scale features f_d^M:

$$f_d^M = \text{SDC}_d\left(\text{BN}\left(C_3\left(f_{ca}^M\right)\right)\right), \quad d = 1, 3, 5, 7, \tag{14}$$

where $C_3(\cdot)$ represents a 3×3 convolutional layer, $\text{SDC}_d(\cdot)$ denotes a smooth dilated convolutional layer with dilation factor d and a kernel size of 3×3. Then, the four features are concatenated and passed through a convolutional layer to obtain the multi-scale pixel-attention weighted feature f_{pa}^M:

$$f_{pa}^M = \sigma\left(C_1\left(\delta\left(\text{BN}\left(C_3\left(\text{concat}\left(f_1^M, f_3^M, f_5^M, f_7^M\right)\right)\right)\right)\right)\right) \otimes f_{ca}^M. \tag{15}$$

Finally, a residual connection combines the input feature with the pixel-attention feature to ensure effective gradient propagation, yielding the output feature f_{out}^M:

$$f_{\text{out}}^M = f_{pa}^M + f_{\text{in}}^M. \tag{16}$$

In parallel, I is fed to CSAM's contrast self-correction branch: a pretrained dehazing model (AOD-Net [12]) generates a pseudo-dehazed image I_{pd} from which a contrast map C_{sa} is computed. We choose AOD-Net as our pseudo-dehazing reference mainly for two reasons. First, it is an ultra-lightweight end-to-end network with negligible computational overhead. Second, it can reliably generate pseudo-dehazed images with basic contrast and object contours preserved, which fully meets the demand of our contrast-guided attention correction.

From the pseudo-dehazed image I_{pd}, we compute the local RMS contrast map C_{sa} with a 5×5 sliding window, whose exact formulation is:

$$C_{sa}(x, y) = \sqrt{\frac{1}{N} \sum_{(i,j) \in \Omega_{5\times5}(x,y)} \left(I_{pd}(i, j) - \mu_\Omega(x, y)\right)^2}, \tag{17}$$

where $\Omega_{5\times5}(x, y)$ is the 5×5 sliding window centred at pixel (x, y), $N = 25$ is the total number of pixels in the window, I_{pd} is the grayscale map of the pseudo-dehazed image and $\mu_\Omega(x, y)$ is the mean grayscale value of pixels in the sliding window. The calculated RMS contrast map is further min-max normalised to the range of $[0, 1]$ per image for subsequent attention correction.

We choose the RMS contrast metric for its parameter-free property, high computational efficiency and universal response to object contours across different haze scenes.

The original attention-weighted features are then locally corrected by C_{sa}:

$$F_{csam} = (1-\alpha)F + \alpha(C_{sa} \otimes F), \tag{18}$$

where F_{csam} is the attention-weighted feature after contrast correction, and α is a learnable parameter initialised to 0.5 that is adaptively adjusted during training and $\otimes$ denotes element-wise multiplication. Through this design, CSAM integrates local contrast correction and multi-scale attention to guide the model's attention towards high-contrast regions that align with object contours and salient structures.

3.4 | Loss Functions

This section adopts joint optimisation losses that integrate detection loss and dehazing loss. This strategy encourages the dehazed images to achieve high perceptual quality while remaining effective for downstream vision tasks.

3.4.1 | *Detection Loss*

The detection loss $\mathcal{L}_{\text{det}}$ adopts the standard three-term objective of YOLOv5s [49] to ensure that dehazed images remain consistent with downstream detection requirements: $\mathcal{L}_{\text{det}} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$, where $\mathcal{L}_{\text{box}}$ is the bounding box regression term, and $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{cls}}$ are binary cross-entropy losses applied to objectness and class predictions. The hyperparameter settings follow the official YOLOv5s [49] configuration where the coefficients λ_{box}, λ_{obj} and λ_{cls} serve as balancing factors among the three loss components and are dynamically scaled depending on the number of classes, image size and the number of detection layers. During training, the YOLOv5s detector remains frozen; $\mathcal{L}_{\text{det}}$ is computed with the frozen detector, and gradients are propagated only through the dehazing network.

3.4.2 | *Dehazing Loss*

The dehazing loss $\mathcal{L}_{\text{dehaze}}$ is composed of two components. The unsupervised part follows [50] and introduces several visual-quality-driven nonreference terms to guide the model without paired supervision. The supervised part adopts the mean absolute error (MAE) between the dehazed output and the ground-truth image, providing pixel-level fidelity when clean references are available.

3.4.2.1 | *Unsupervised Loss.* The unsupervised part of the dehazing loss follows [50]. It is defined as $\mathcal{L}_{\text{unsup}} = 0.008\mathcal{L}_{\text{dcp}} + 3\mathcal{L}_{\text{con}} + 0.05\mathcal{L}_{\text{sat}} + 0.0008\mathcal{L}_{\text{sharp}}$, where $\mathcal{L}_{\text{dcp}}$, $\mathcal{L}_{\text{con}}$, $\mathcal{L}_{\text{sat}}$ and $\mathcal{L}_{\text{sharp}}$ denote the dark channel loss, contrast loss, saturation loss and edge sharpness loss terms in [50]. These losses encourage the network to suppress residual haze, enhance contrast, restore colour saturation and preserve structural details. The hyperparameter settings follow [50]. We also tested different parameter configurations and observed that the original settings remained robust in our framework. Therefore, these configurations are adopted in all experiments.

3.4.2.2 | *Mean Absolute Error Loss.* The mean -absolute error (MAE) loss $\mathcal{L}_{\text{mae}}$ measures the difference between the generated dehazed image and the ground-truth image, guiding the network to optimise during training so that the output approaches the haze-free image:

$$\mathcal{L}_{\text{mae}} = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\|J(i,j) - GT(i,j)\|_1, \tag{19}$$

where $J(i,j)$ and $GT(i,j)$ are the pixel values of the dehazed image and the ground-truth image at (i,j), respectively. H and W are the image height and width.

3.4.3 | *Total Loss*

The total loss is a linear combination of the above losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{unsup}}\mathcal{L}_{\text{unsup}} + \lambda_{\text{mae}}\mathcal{L}_{\text{mae}}, \tag{20}$$

here λ_{det}, λ_{unsup} and λ_{mae} are weighting coefficients that balance the contributions of the three terms. These coefficients are determined through a grid search over multiple parameter configurations and are finally set to $\lambda_{\text{det}} = 0.4$, $\lambda_{\text{unsup}} = 0.4$ and $\lambda_{\text{mae}} = 0.6$, which provide the best trade-off between dehazing quality and detection accuracy.

4 | Experiments and Results

This section presents both qualitative and quantitative comparisons with state-of-the-art dehazing approaches on synthetic and real-world datasets to evaluate the effectiveness of the proposed method. Specifically, the performance is benchmarked against a broad range of representative baselines, including traditional prior-based methods (FVR [51], DCP [8]), early CNN-based models (DehazeNet [11], AOD-Net [12], and EPDN [25]) and more recent weakly-supervised learning methods such as RefineDNet [32], YOLY [37], DMT [33], DNMGDT [52], SDA-GAN [7] and DNDM [6]. In addition, ablation studies are conducted to assess the contribution of core modules and loss functions within the proposed framework.

4.1 | Experimental Settings

4.1.1 | *Datasets and Evaluation Metrics*

The training set consists of 3000 synthetic hazy images from Foggy Cityscapes (FC) [23] with three fog densities ($\beta \in \{0.005, 0.01, 0.02\}$). In addition, 1000 real hazy images from RTTS [22] are included for unsupervised training. Evaluation is conducted on multiple benchmarks: SOTS (500 outdoor synthetic and 500 indoor synthetic) and HSTS (10 outdoor synthetic) [22]. For detection experiments, 500 Foggy Cityscapes images that are not part of the training set are used. To assess real-world applicability, 500 real hazy images from RTTS, also excluded from training, are adopted for testing. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are reported on synthetic datasets with ground truth. Qualitative comparisons are provided on real-world datasets without ground truth. Mean average precision at IoU threshold 0.5 (mAP@0.5) is used to evaluate detection performance.

4.1.2 | *Implementation Details*

A pre-trained YOLOv5s [49], widely adopted in prior works [53, 54], is used as the detector in all experiments. During training, the detector is frozen, and the detection loss is computed by forwarding dehazed images through it with gradients propagated only to the dehazing network. The same detector is used for evaluation with default inference settings (confidence threshold 0.25 and NMS IoU as in the official release) without test-time augmentation or multi-scale testing. Qualitative results are presented by retaining only the person and car classes for visualisation. The hyperparameters in the total loss are determined through grid search across multiple configurations. The final weights are set to $\lambda_{\text{det}} = 0.4, \lambda_{\text{unsup}} = 0.4$ and $\lambda_{\text{mae}} = 0.6$, which achieve the best balance between dehazing quality and detection accuracy. All other coefficients in the detection and unsupervised losses follow the default settings of YOLOv5s [49] and Yang et al. [50], respectively. We also experimented with different parameter configurations but found that the default settings in these works remained stable and effective within our framework. The Adam optimiser is adopted with exponential decay rates $\beta_1 = 0.9, \beta_2 = 0.999$. The initial learning rate is set to 0.0002 and dynamically adjusted by a cosine annealing schedule during training. The batch size is fixed at 16. The network is trained for 200 epochs, which takes about three days on a single GPU.

4.1.3 | *Training Protocol*

In this subsection, we systematically elaborate the complete training workflow of our Dehaze4Detection framework.

4.1.3.1 | *State of Pre-Trained Components.* Our framework relies on three pre-trained components, all of which are fully frozen during the entire training and inference process, with no parameter updates. They are only used to provide supervision signals and prior information: the frozen YOLOv5s detector is used to compute the detection loss $\mathcal{L}_{\text{det}}$ for task alignment; the frozen PSPNet segmentation model is used to generate pixel-level semantic maps for the SMFM module; and the frozen AOD-Net dehazing model is used to generate pseudo-dehazed images for contrast prior extraction in the CSAM module. All three pre-trained components are applied consistently across synthetic and real-world data.

4.1.3.2 | *Supervision Loss Allocation.* We adopt different supervision strategies for synthetic and real-world data according to the availability of paired clean ground truth. For the 3000 synthetic hazy images from Foggy Cityscapes with paired clean ground truth, we apply the full loss function, including the MAE reconstruction loss $\mathcal{L}_{mae}$, unsupervised dehazing loss $\mathcal{L}_{unsup}$ and detection loss $\mathcal{L}_{\text{det}}$. The reconstruction loss is only applied to the final dehazed output of the network with no additional supervision constraints at intermediate scales. For the 1000 real hazy images from RTTS without paired clean ground truth, we only apply the detection loss $\mathcal{L}_{\text{det}}$ and unsupervised dehazing loss $\mathcal{L}_{unsup}$ without using the MAE reconstruction loss.

4.1.3.3 | *Training Pipeline.* We adopt an intra-batch mixing strategy for synthetic and real data. In each training batch, the ratio of synthetic images to real images is fixed at 3:1 consistent with the sample size ratio of the training set. Synthetic and real samples are mixed and processed in the same training batch. During forward propagation, all samples in the same batch complete forward propagation simultaneously: the input hazy image is fed into the CSAM module for contrast-guided feature refinement, then into the SMFM module for semantic-aware multi-scale fusion, and finally generates the dehazed output. The dehazed output is then fed into the frozen YOLOv5s detector to compute the detection loss. During backward propagation, the losses of all samples in the batch are aggregated into the total loss for unified backpropagation. Gradients are only propagated through the dehazing backbone network (CSAM + SMFM) with no gradient updates to the three frozen pre-trained components.

4.2 | Experiments on Synthetic Datasets and Analysis

To begin with, Table 1 summarises the quantitative dehazing results on three widely used synthetic datasets: SOTS, HSTS and Foggy Cityscapes (FC). The proposed method provides balanced improvements across datasets. On SOTS, it achieves the highest SSIM (0.931) with the second-best PSNR (26.07 dB, 0.12 dB lower than DMT), indicating stronger structural fidelity at comparable distortion. On the more challenging HSTS, it attains the best scores on both PSNR (29.05 dB) and SSIM (0.960), outperforming strong competitors such as DNDM (28.53 dB/0.933), which indicates superior robustness under dense haze. On Foggy Cityscapes (FC), it also yields the highest PSNR (20.04 dB) and the second-best SSIM (0.885 vs. 0.887 for DNDM). Overall, the consistent gains across the three benchmarks suggest that the proposed method improves both restoration fidelity and structure preservation across scene types and haze levels. Complementing the quantitative evaluation, the dehazing performance of the proposed method is further analysed through visual comparisons of dehazed images obtained by different methods on the SOTS and HSTS test sets as shown in Figures 6 and 7. From Figures 6b and 7b, it can be observed that DCP introduces obvious artefacts in the dehazed images, and the contrast and colour of the images show significant distortion, especially in the sky regions. From Figures 6c,d and 7c,d, it can be observed that DehazeNet and AOD-Net rely on fixed scattering-model parameters and cannot adaptively handle complex hazy scenes resulting in residual haze in the dehazed images. EPDN yields over-enhanced colour saturation after dehazing, and locally dark phenomena appear in the images, for example, the trees and buildings in Figure 6e; at the same time, due to over-strengthening of colour information, the dehazed images show obvious colour distortions, for example, the sky region in Figure 7e. From Figures 6d–f and 7d–f, it can also be seen that RefineDNet and AOD-Net both rely on physical models, and RefineDNet uses prior information for preprocessing, thus exhibiting defects similar to DCP with distortions in the sky region. DMT removes most haze but leaves slight residual haze in regions with complex backgrounds, for example, the tree area in the first image and the crowd area in the second image of Figure 6g. As shown in Figure 6h, the results of DNMGDT on SOTS still contain residual haze particularly in regions with complex backgrounds. SDA-GAN exhibits obvious colour shifts after dehazing especially in low-illumination or low-contrast regions, such as the building region in the third image of Figure 6i and the sky region in the third image of Figure 7i. The images produced by DNDM also

TABLE 1 | Quantitative dehazing results on SOTS, HSTS and FC.

Method	SOTS		HSTS		FC	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
FVR (2009)	11.39	0.590	13.2	0.485	14.35	0.592
DCP (2010)	14.25	0.716	15.38	0.655	16.73	0.634
DehazeNet (2016)	18.52	0.711	20.57	0.819	17.21	0.629
AOD-Net (2017)	17.94	0.671	19.34	0.792	16.47	0.617
EPDN (2019)	20.22	0.782	18.47	0.775	16.55	0.649
RefineDNet (2021)	21.17	0.856	20.12	0.815	18.15	0.781
YOLY (2021)	23.80	0.927	22.06	0.901	18.27	0.792
DMT (2021)	**26.19**	0.911	28.14	0.939	19.55	0.876
SDA-GAN (2022)	19.53	0.802	19.77	0.790	15.41	0.714
DNDM (2023)	24.50	0.898	28.53	0.933	19.23	**0.887**
DNMGDT (2025)	24.67	0.918	26.29	0.914	18.93	0.834
Ours	26.07	**0.931**	**29.05**	**0.960**	**20.04**	0.885

Note: Red/blue/green indicate the best, second-best and third-best results, respectively.

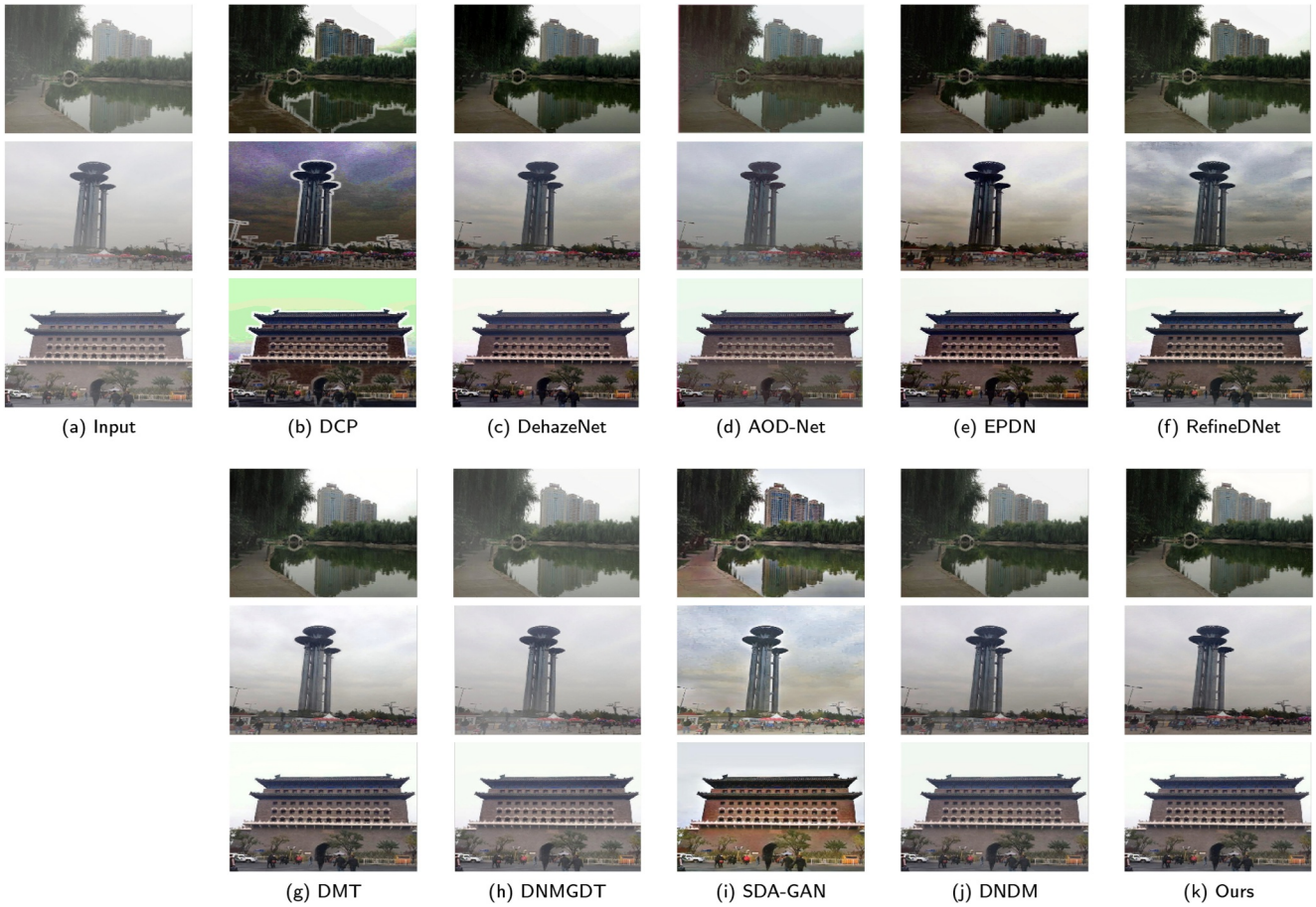


FIGURE 6 | Qualitative dehazing results on the SOTS dataset. Our method restores image contrast and edge-texture details while maintaining high-quality haze removal with better structural fidelity than compared methods.

exhibit localised haze residuals and blurred fine details particularly in the tree regions of the first image in Figure 6j and the second image in Figure 7j. Compared with other methods, the proposed method restores image contrast and edge-texture details while maintaining high-quality haze removal, yielding visually consistent results with better structural fidelity.

Beyond the above analysis, Table 2 reports quantitative dehaze–then–detect results on Foggy Cityscapes (FC) with a fixed YOLOv5s under three haze densities ($\beta \in \{0.005, 0.01, 0.02\}$). The proposed method achieves the best average mAP@0.5 of 57.3% surpassing the second-best DNMGDT by +1.2% (57.3% vs. 56.1%). Specifically, under light haze ($\beta = 0.005$), it attains the

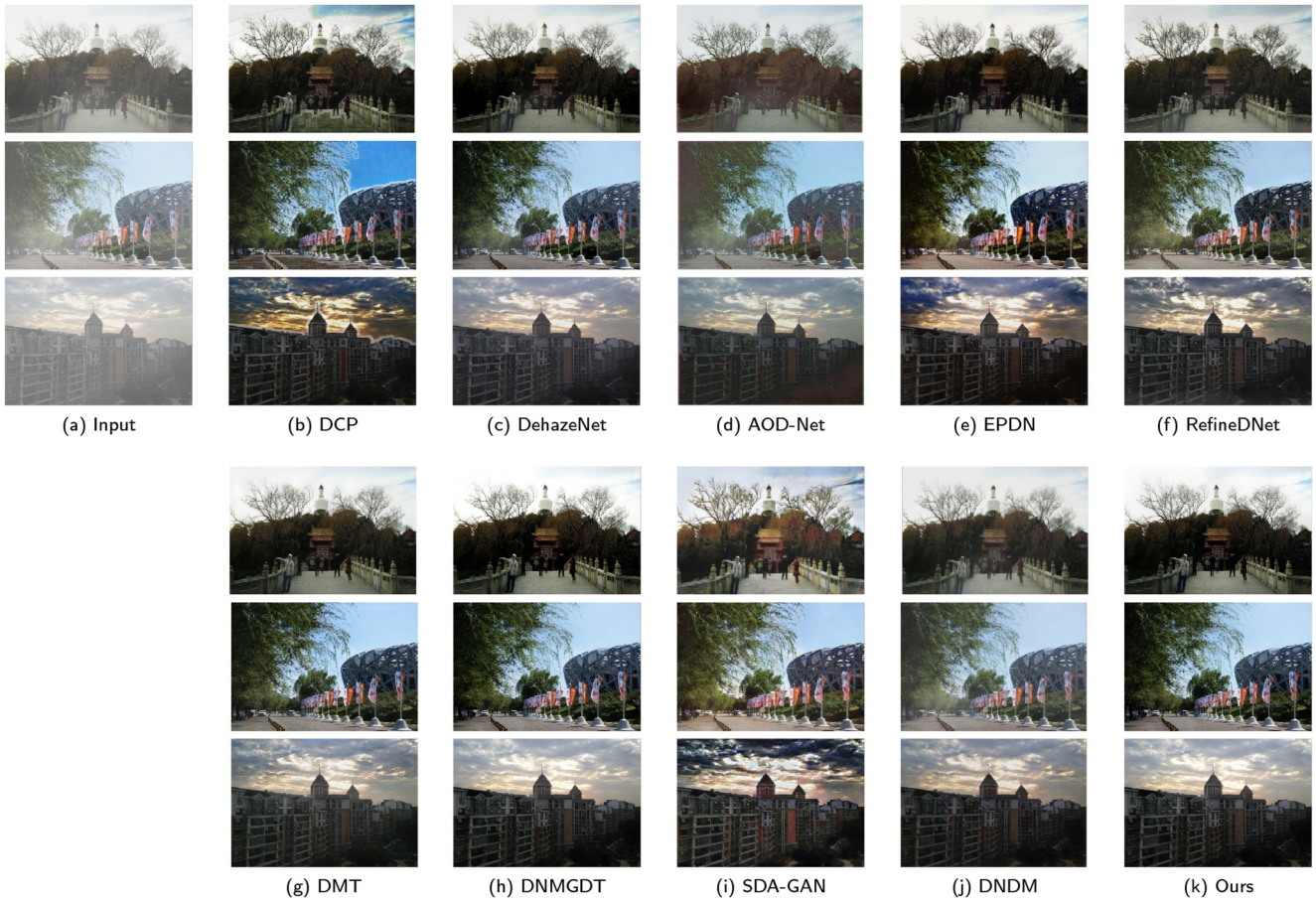


FIGURE 7 | Qualitative dehazing results on the HSTS dataset. Our method recovers contrast and edge textures while maintaining colour consistency, delivering clearer object contours and better structural fidelity.

TABLE 2 | Quantitative dehaze–then–detect results on Foggy Cityscapes (FC) with YOLOv5s (mAP@0.5↑).

Type	Method	Haze density			Average mAP@0.5 ↑
		0.005	0.01	0.02	
Hazy image	YOLOv5s	60.5	55.3	44.6	53.5
Prior-based	FVR + YOLOv5s	60.3	55.2	44.7	53.4
	DCP + YOLOv5s	60.8	55.4	45.3	53.8
Supervised	DehazeNet + YOLOv5s	57.2	50.1	41.7	49.7
	AOD-Net + YOLOv5s	56.5	51.7	41.1	49.8
	EPDN + YOLOv5s	63.4	56.3	43.6	54.4
Weakly supervised	RefineDNet + YOLOv5s	62.9	57.4	47.3	55.9
	DMT + YOLOv5s	63.1	56.8	48.2	56.0
	DNMGDT + YOLOv5s	63.1	56.6	48.5	56.1
	SDA-GAN + YOLOv5s	57.6	54.5	46.9	53.0
	DNDM + YOLOv5s	63.3	56.3	44.0	54.5
	Ours + YOLOv5s	**65.1**	**57.9**	**48.9**	**57.3**

Note: Red/blue/green indicate the best, second-best and third-best results, respectively.

highest mAP@0.5 of 65.1%, outperforming EPDN (63.4%) and DNDM (63.3%). At medium haze ($\beta = 0.01$), it delivers the best mAP@0.5 of 57.9% leading the second-best RefineDNet by +0.5% (57.9% vs. 57.4%). Even under dense haze ($\beta = 0.02$), where detection becomes more challenging, it ranks first (48.9%), outperforming DNMGDT (48.5%) and DMT (48.2%), indicating consistent robustness across haze densities. These results confirm that the proposed method can not only restore haze-degraded images but also enhance task-relevant features to support robust detection under different haze conditions. Complementing the quantitative evaluation, Figure 8 shows qualitative dehaze–then–detect comparisons on the Foggy Cityscapes test

set. From Figure 8b,c, both DMT and DNMGDT yield darker dehazed images and exhibit false positives; for instance, nonexistent cars are erroneously detected in the second image of Figure 8b,c. SDA-GAN shows obvious colour shifts after dehazing, and there are more erroneous detections (e.g., the first image of Figure 8d misrecognises a building as a person and a road as a car). The dehazed images generated by DNDM also lead to false positives in dense haze regions (e.g., extra detections of nonexistent cars in the third image of Figure 8e). Compared with other methods, the proposed method can remove haze while enhancing target-region features, restoring image sharpness and contrast, producing stable detection with clearer contours and more reliable confidence.

4.3 | Experiments on Real-World Datasets and Results Analysis

This section evaluates the dehazing performance of the proposed method on real hazy images by randomly selecting three dehazed images from the RTTS test set for visual comparisons, as shown in Figure 9. DMT estimates atmospheric light and transmission in steps; this process can cause cumulative estimation errors, leading to distortions in the dehazed images, such as the sky region in the first image of Figure 9b. As shown in Figure 9c, the dehazed images by DNMGDT are overall blurry, and there remains residual haze. SDA-GAN alleviates the domain shift problem by aligning the source and target domains. However, due to incomplete domain alignment, the model may fail to accurately capture the colour characteristics and haze characteristics of the target-domain images, leading to residual haze and over-saturated colours (e.g., the person region in the first image of Figure 9d). From Figure 9e, DNDM fails to generalise well to real-world dehazing resulting in residual haze and insufficient contrast enhancement. The proposed method can effectively adapt to real-scene haze removal, restore texture features and image contrast, and the dehazed images possess realism and naturalness.

Beyond the above analysis, Table 3 summarises the dehaze–then–detect results on the real-world RTTS test set using a fixed YOLOv5s detector (mAP@0.5). The proposed method achieves the best overall score of 63.7%, surpassing the strongest baseline DNMGDT (61.6%) by +2.1% and the hazy-image detector (56.8%) by +6.9%. Compared with other recent methods, including DMT (60.4%) and SDA-GAN (59.8%), our method exhibits clear advantages. Notably, DNDM shows a performance drop to 53.0%, indicating its sensitivity to domain shift and limited generalisation under real hazy conditions.

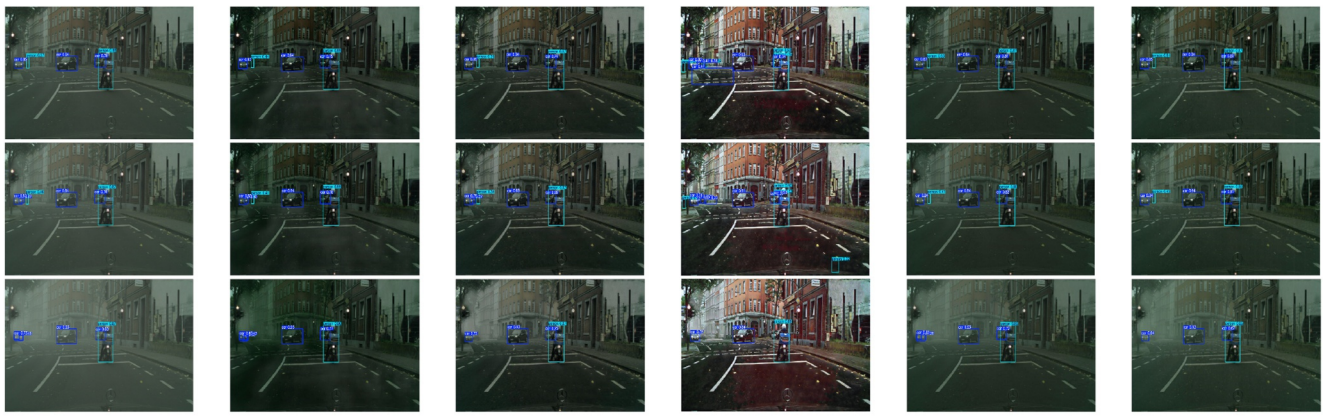


FIGURE 8 | Qualitative dehaze–then–detect comparisons on Foggy Cityscapes with YOLOv5s under three haze densities (β = 0.005, 0.01 and 0.02). Our method yields clearer contours and more stable detections than competing methods.



FIGURE 9 | Qualitative dehazing comparisons on real hazy RTTS images. Our method better restores image textures, contrast and natural colours.

These results highlight the robustness and adaptability of the proposed framework in real-world scenarios. Complementing the quantitative evaluation, this section provides a qualitative analysis of detection performance by randomly selecting three detection results from the real hazy image dataset for visual comparison, as shown in Figure 10.

As shown in Figure 10b, DMT effectively increases overall contrast, but slight residual haze remains in target regions, leading to false and missed detections (e.g., a person is falsely detected near the car wheel in the first image, and multiple persons are missed in the second image). As shown in Figure 10c, DNMGDT leaves residual haze that blurs target textures and causes missed detections (e.g., multiple persons are missed in the second image). As shown in Figure 10d, the dehazed images produced by SDA-GAN exhibit severe colour over-saturation, which leads to both false and missed detections (e.g., part of a car is misrecognised as a person in the first image, whereas in the second image, some persons are missed and a fence is misrecognised as a person). In Figure 10e, DNDM retains severe residual haze, which reduces detection confidence and causes missed detections (e.g., multiple persons are not recognised in the second image, and some cars are not correctly recognised in the third image). As shown in Figure 10f, the proposed method produces natural and realistic contrast, preserves object edge details and yields more accurate and reliable detection.

TABLE 3 | Quantitative dehaze–then–detect results on RTTS with YOLOv5s (mAP@0.5↑).

Methods	mAP@0.5 ↑
YOLOv5s	56.8
DMT + YOLOv5s	60.4
DNMGDT + YOLOv5s	61.6
SDA-GAN + YOLOv5s	59.8
DNDM + YOLOv5s	53.0
Ours	**63.7**

Note: Bold indicates the best result.

4.4 | Ablation Studies

We conduct comprehensive ablation studies on the Foggy Cityscapes dataset to verify the individual and combined contributions of each core component of our framework. For fair comparison, all ablation settings keep the overall network depth and parameter count basically consistent with the full framework. The specific alternative structures for ablation are as follows:

1. wo-SMFM (Without SMFM Module): We replace all SMFB modules with RDB blocks of the same parameter count, removing only the semantic guidance branch and the corresponding channel attention fusion module, while keeping the rest of the network unchanged.
2. wo-CSAM (Without CSAM Module): We replace the complete CSAM structure with standard convolutional blocks (CB) and MAB blocks of the same parameter count, removing only the contrast self-correction branch, while keeping the rest of the network unchanged.
3. wo-$\mathcal{L}_{\text{det}}$ (Without Detection Loss): We remove the detection loss from the total loss function, while retaining the complete SMFM and CSAM modules and the rest of the loss terms.

Table 4 presents the quantitative ablation results on the Foggy Cityscapes dataset in terms of PSNR, SSIM and mAP@0.5. The full model achieves the best overall performance, attaining a 20.04 dB PSNR, 0.885 SSIM and 57.3% mAP@0.5. Removing the SMFM results in the largest performance drop, reducing mAP from 57.3% to 54.9%, PSNR from 20.04 to 18.12 dB and SSIM

TABLE 4 | Quantitative ablation results on the FC dataset.

Model	SMFM	CSAM	$\mathcal{L}_{\text{det}}$	PSNR	SSIM	mAP@0.5
wo-SMFM		✓	✓	18.12	0.827	54.9
wo-CSAM	✓		✓	19.61	0.860	56.2
wo-$\mathcal{L}_{\text{det}}$	✓	✓		19.96	0.881	56.7
Ours	✓	✓	✓	20.04	0.885	57.3

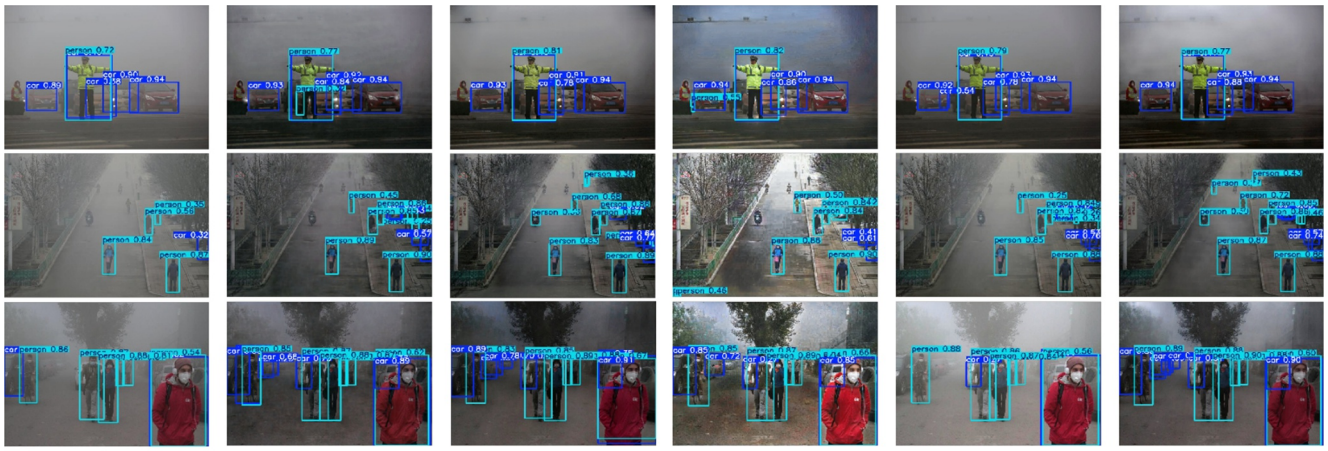


FIGURE 10 | Qualitative dehaze–then–detect comparisons on RTTS with YOLOv5s. Our method preserves clearer object boundaries and achieves more reliable detections under real haze conditions.



FIGURE 11 | Qualitative ablation results on synthetic Foggy Cityscapes images (top) and real RTTS images (bottom). The full model produces clearer restoration and more reliable detections than the variants without SMFM, CSAM or detection loss.

from 0.885 to 0.827. This confirms the importance of embedding pixel-level semantic knowledge and multi-scale features throughout the reconstruction for enhancing task-relevant regions. Without CSAM, SSIM falls to 0.860 and mAP to 56.2%, indicating that leveraging contrast maps from pseudo-dehazed images helps refine attention towards object contours and salient structures, thus improving target–background separability. When the detection loss is removed, the model still yields relatively high PSNR (19.96 dB) and SSIM (0.881), but mAP drops from 57.3% to 56.7%, suggesting that without explicit task-level supervision, the restored features are less aligned with the detection task. These results collectively demonstrate the complementary contributions of SMFM, CSAM and the detection loss in producing visually faithful and detection-friendly dehazed images. Complementing the quantitative evaluation, the qualitative ablation results on synthetic and real images are shown in Figure 11. The top row and bottom row correspond to the test results on synthetic hazy images from the Foggy Cityscapes dataset and real hazy images from the RTTS dataset, respectively. From Figure 11a, the results show that without SMFM, target regions tend to retain haze while background areas are over-enhanced and noisy due to the lack of pixel-level semantic guidance. This degrades the dehazing quality and leads to missed detections (e.g., in the second image of Figure 11a, the car is not correctly recognised). From Figure 11b, the results show that without CSAM, removing contrast self-correction leaves residual haze along object contours and blurs target edges, which weakens object-background discriminability and can make the model ignore key structures (e.g., residual haze remains at human contours in the second image of Figure 11b), thereby reducing detection accuracy. From Figure 11c, it can be observed that without the detection loss, the alignment between restoration and the detection objective becomes weaker and is mainly reflected in lower detection confidence. In summary, the proposed SMFM, CSAM and the detection loss function play important and complementary roles.

5 | Conclusion

This work presents a task-aligned weakly supervised dehazing framework, termed Dehaze4Detection, which explicitly aligns image restoration with object detection. The framework integrates a Semantic-Aware Multi-Scale Fusion Module that embeds pixel-level semantic knowledge throughout the reconstruction process, enabling selective enhancement of task-relevant regions. It further introduces a Contrast Self-Correction Attention Module that leverages contrast maps from pseudo-dehazed images to refine attention towards object contours and salient structures, reducing artefacts and improving structural clarity. The training objective jointly optimises task-aligned dehazing and detection losses so that the restored images are visually clear and semantically aligned with downstream detection requirements. Experimental results on RESIDE (SOTS, HSTS, RTTS) and Foggy Cityscapes demonstrate that the proposed method achieves competitive performance in both visual quality and detection accuracy. Future work will further investigate the influence of semantic priors on task-aligned dehazing under challenging haze conditions. The dependence of contrast-guided correction on pseudo-dehazed images also deserves deeper exploration together with more robust strategies to alleviate the potential influence of inaccurate prior information. These directions are expected to further improve the stability and adaptability of the proposed framework in complex real-world scenarios.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant 2022ZD0160400) and National Science Fund for Distinguished Young Scholars (61925602). We gratefully acknowledge the support from Shanghai Artificial Intelligence Laboratory.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data available on request from the authors.

References

1. Q. Zhang, J. Xiao, C. Tian, J. C.-W. Lin, and S. Zhang, "A Robust Deformed Convolutional Neural Network (CNN) for Image Denoising," *CAAI Transactions on Intelligence Technology* 8, no. 2 (2023): 331–342, https://doi.org/10.1049/cit2.12110.

2. C. Tian, X. Zhang, Qi Zhang, M. Yang, and Z. Ju, "Image Super-Resolution via Dynamic Network," *CAAI transactions on intelligence technology* 9, no. 4 (2024): 837–849, https://doi.org/10.1049/cit2.12297.

3. B. Goyal, A. Dogra, D. C. Lepcha, et al., "Recent Advances in Image Dehazing: Formal Analysis to Automated Approaches," *Information Fusion* 104 (2024): 102151, https://doi.org/10.1016/j.inffus.2023.102151.

4. A. Murtaza, U. Khairuddin, A.'A. Mohd Faudzi, K. Hamamoto, Y. Fang, and Z. Omar, "WaveLiteDehaze-Network: A Low-Parameter Wavelet-Based Method for Real-Time Dehazing," *CAAI Transactions on Intelligence Technology* 10, no. 4 (2025): 1033–1048, https://doi.org/10.1049/cit2.70011.

5. Z. He, P. Wang, Y. Zou, et al., "Dadigan: Aa Dual Attention Blocks-Based Disentangled Iterative Generative Adversarial Network for Cloud and Shadow Removal on SAR and Optical Images," *Information Fusion* 125 (2026): 103487, https://doi.org/10.1016/j.inffus.2025.103487.

6. T. Jia, J. Li, Li Zhuo, and T. Yu, "Semi-Supervised Single-Image Dehazing Network via Disentangled Meta-Knowledge," *IEEE Transactions on Multimedia* 26 (2023): 2634–2647, https://doi.org/10.1109/tmm.2023.3301273.

7. Yu Dong, Y. Li, Q. Dong, He Zhang, and S. Chen, "Semi-Supervised Domain Alignment Learning for Single Image Dehazing," *IEEE Transactions on Cybernetics* 53, no. 11 (2022): 7238–7250, https://doi.org/10.1109/tcyb.2022.3221544.

8. K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, no. 12 (2010): 2341–2353, https://doi.org/10.1109/TPAMI.2010.168.

9. Q. Zhu, J. Mai, and L. Shao, "A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior," *IEEE Transactions on Image Processing* 24, no. 11 (2015): 3522–3533, https://doi.org/10.1109/tip.2015.2446191.

10. D. Berman and S. Avidan, "Non-Local Image Dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1674–1682.

11. B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An End-to-End System for Single Image Haze Removal," *IEEE Transactions on Image Processing* 25, no. 11 (2016): 5187–5198, https://doi.org/10.1109/tip.2016.2598681.

12. B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-Net: All-In-One Dehazing Network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), 4770–4778.

13. X. Yin, G. Tu, and Q. Chen, "Multiscale Depth Fusion With Contextual Hybrid Enhancement Network for Image Dehazing," *IEEE Transactions on Instrumentation and Measurement* 72 (2023): 1–12, https://doi.org/10.1109/tim.2023.3318748.

14. Z. Lihe, J. He, Q. Yuan, X. Jin, Y. Xiao, and L. Zhang, "Phdnet: A Novel Physic-Aware Dehazing Network for Remote Sensing Images," *Information Fusion* 106 (2024): 102277, https://doi.org/10.1016/j.inffus.2024.102277.

15. Y. Song, Z. He, H. Qian, and X. Du, "Vision Transformers for Single Image Dehazing," *IEEE Transactions on Image Processing* 32 (2023): 1927–1941, https://doi.org/10.1109/tip.2023.3256763.

16. Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. L. Uformer, "A General u-Shaped Transformer for Image Restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 17683–17693.

17. Y. Wang, J. Xiong, X. Yan, and M. Wei, "Uscformer: Unified Transformer With Semantically Contrastive Learning for Image Dehazing," *IEEE Transactions on Intelligent Transportation Systems* 24, no. 10 (2023): 11321–11333, https://doi.org/10.1109/tits.2023.3277709.

18. W. Yu, Q. Wang, C. Liu, L. Dong, and Q. Hu, "Coe: Chain-of-Explanation via Automatic Visual Concept Circuit Description and Polysemanticity Quantification," in *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), 4364–4374.

19. C. Tian, T. Cheng, Y. Ma, Z. Qi, B. Zhang, and D. Zhang, "A Heterogeneous CNN With Transformers for Image Denoising," *IEEE Transactions on Consumer Electronics* 72, no. 2 (2026): 2658–2667, https://doi.org/10.1109/tce.2026.3665312.

20. Y. Fan, Y. Wang, M. Wei, F. L. Wang, and H. Xie, "Friendnet: Detection-Friendly Dehazing Network," preprint, arXiv:2403.04443 (2024), https://arxiv.org/html/2403.04443v1.

21. Y. Wan, J. Li, L. Lin, Q. Yuan, and H. Shen, "Collaboration of Dehazing and Object Detection Tasks: A Multi-Task Learning Framework for Foggy Image," *IEEE Transactions on Geoscience and Remote Sensing* (2025), https://doi.org/10.1109/TGRS.2025.3548283.

22. B. Li, W. Ren, D. Fu, et al., "Benchmarking Single-Image Dehazing and Beyond," *IEEE Transactions on Image Processing* 28, no. 1 (2018): 492–505, https://doi.org/10.1109/tip.2018.2867951.

23. C. Sakaridis, D. Dai, and L. Van Gool, "Semantic Foggy Scene Understanding With Synthetic Data," *International Journal of Computer Vision* 126, no. 9 (2018): 973–992, https://doi.org/10.1007/s11263-018-1072-8.

24. W. Ren, Si Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single Image Dehazing via Multi-Scale Convolutional Neural Networks," in *European Conference on Computer Vision (ECCV)* (Springer, 2016), 154–169.

25. Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix Dehazing Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 8160–8168.

26. D. Chen, M. He, Q. Fan, et al., "Gated Context Aggregation Network for Image Dehazing and Deraining," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2019), 1375–1383.

27. X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-Based Multi-Scale Network for Image Dehazing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 7314–7323.

28. H. Dong, J. Pan, L. Xiang, et al., "Multi-Scale Boosted Dehazing Network With Dense Feature Fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 2157–2167.

29. Xu Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-Net: Feature Fusion Attention Network for Single Image Dehazing," *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 7 (2020): 11908–11915, https://doi.org/10.1609/aaai.v34i07.6865.

30. H. Wu, Y. Qu, S. Lin, et al., "Contrastive Learning for Compact Single Image Dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10551–10560.

31. C.-Le Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image Dehazing Transformer With Transmission-Aware 3d Position Embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 5812–5820.

32. S. Zhao, L. Zhang, Y. Shen, and Y. Zhou, "Refinednet: A Weakly Supervised Refinement Framework for Single Image Dehazing," *IEEE Transactions on Image Processing* 30 (2021): 3391–3404, https://doi.org/10.1109/tip.2021.3060873.

33. Ye Liu, L. Zhu, S. Pei, et al., "From Synthetic to Real: Image Dehazing Collaborating With Unlabeled Real Data," in *Proceedings of the ACM International Conference on Multimedia (MM)* (2021), 50–58.

34. A. Golts, D. Freedman, and M. Elad, "Unsupervised Single Image Dehazing Using Dark Channel Prior Loss," *IEEE Transactions on Image*

Processing 29 (2019): 2692–2701, https://doi.org/10.1109/tip.2019.2952032.

35. B. Li, Y. Gou, J. Z. Liu, H. Zhu, J. T. Zhou, and Xi Peng, "Zero-Shot Image Dehazing," *IEEE Transactions on Image Processing* 29 (2020): 8457–8466, https://doi.org/10.1109/tip.2020.3016134.

36. D. Engin, A. Genç, and H. K. Ekenel, "Cycle-Dehaze: Enhanced Cyclegan for Single Image Dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018).

37. B. Li, Y. Gou, S. Gu, J. Z. Liu, J. T. Zhou, and X. Peng, "You Only Look Yourself: Unsupervised and Untrained Single Image Dehazing Neural Network," *International Journal of Computer Vision* 129, no. 5 (2021): 1754–1767, https://doi.org/10.1007/s11263-021-01431-5.

38. Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain Adaptation for Image Dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 2808–2817.

39. H. Yu, J. Huang, Y. Liu, Q. Zhu, M. Zhou, and F. Zhao, "Source-Free Domain Adaptation for Real-World Image Dehazing," in *Proceedings of the ACM International Conference on Multimedia (MM)* (2022), 6645–6654.

40. Y. Wang, X. Yan, K. Zhang, et al., "Togethernet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning," *Computer Graphics Forum* 41, no. 7 (2022): 465–476, https://doi.org/10.1111/cgf.14692.

41. C. Li, H. Zhou, Y. Liu, et al., "Detection-Friendly Dehazing: Object Detection in Real-World Hazy Scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 7 (2023): 8284–8295, https://doi.org/10.1109/TPAMI.2023.3234976.

42. Y. Zhu, T. Wang, X. Fu, et al., "Learning Weather-General and Weather-Specific Features for Image Restoration Under Multiple Adverse Weather Conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 21747–21758.

43. Z. Zhang, L. Zhao, Y. Liu, S. Zhang, and J. Yang, "Unified Density-Aware Image Dehazing and Object Detection in Real-World Hazy Scenes," in *Proceedings of the Asian Conference on Computer Vision (ACCV)* (2020), 119–135.

44. W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-Adaptive Yolo for Object Detection in Adverse Weather Conditions," *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 2 (2022): 1792–1800, https://doi.org/10.1609/aaai.v36i2.20072.

45. S. Kalwar, D. Patel, A. Aanegola, K. Reddy Konda, S. Garg, and K. M. Krishna, "Gdip: Gated Differentiable Image Processing for Object Detection in Adverse Conditions," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), 7083–7089.

46. H. Zhang, L. Xiao, X. Cao, and H. Foroosh, "Multiple Adverse Weather Conditions Adaptation for Object Detection via Causal Intervention," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 3 (2022): 1742–1756, https://doi.org/10.1109/tpami.2022.3166765.

47. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2881–2890.

48. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 2472–2481.

49. G. Jocher, A. Stoken, J. Borovec, et al., "ultralytics/yolov5: v3.0," *Zenodo* [Online] (2020), https://doi.org/10.5281/zenodo.3983579.

50. A. Yang, Y. Liu, J. Wang, et al., "Visual-Quality-Driven Unsupervised Image Dehazing," *Neural Networks* 167 (2023): 1–9, https://doi.org/10.1016/j.neunet.2023.08.010.

51. J.-P. Tarel and N. Hautiere, "Fast Visibility Restoration From a Single Color or Gray Level Image," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2009), 2201–2208.

52. Y. Su, N. Wang, Z. Cui, Y. Cai, C. He, and A. Li, "Real Scene Single Image Dehazing Network With Multi-Prior Guidance and Domain Transfer," *IEEE Transactions on Multimedia* (2025).

53. H. Wang, Y. Xu, Y. He, et al., "Yolov5-fog: A Multiobjective Visual Detection Algorithm for Fog Driving Scenes Based on Improved yolov5," *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 1–12, https://doi.org/10.1109/tim.2022.3196954.

54. Z.-D. Zhang, B. Zhang, Z.-C. Lan, et al., "Finet: An Insulator Dataset and Detection Benchmark Based on Synthetic Fog and Improved yolov5," *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 1–8, https://doi.org/10.1109/tim.2022.3194909.