Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Contrast and Visibility Enhancement of Weather-Degraded Images Using Dual-Branch CNN and Transformer with Perceptual Loss for ADAS

Anmol Jain[1], Veerendra Yadav[2], Harsh Khatter[3]

[1,2]Department of Computer Science and Engineering, Noida International University, Gautam Budh Nagar, Uttar Pradesh, India.
[3]Department of Computer Science and Information Technology, Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh, India.

E-mail: [1]anmol.suvrat@gmail.com, [2]vyadav1@ce.iitr.ac.in, [3]harsh.khatter@kiet.edu

Orcid ID: [1]0009-0009-7609-8482, [2]0000-0002-8679-132X, [3]0000-0002-4758-2971

Abstract

Weather can be either poor or good. Poor weather, includING fog, hazE, rain, or low light, can cause dramatic degradation of image perception in road-level situations, leading to with significant performance loss in camera-based Advanced Driver-Assistance Systems (ADAS), Although traditional improvement techniques relying on Convolutional Networks (CNNs) cannot effectively preserve global context in image appearance improvement, techniques using transformers show high computational costs. This restricts their application as real-time system efficiency becomes critically important. In this paper, we propose a solution using the Dual-Branch CNN Transformer, which uniformly utilizes localized spatial features extraction together with global semantic modeling using parallel experience sharing of Convolutional Networks and Self Attention Mechanisms. An adaptive gated fusion module integrates these complementary local and global representations through learnable spatial weighting, while perceptual-loss-guided optimization emphasizes texture fidelity, structural consistency, and visual realism. The model was tested on real-world driving image datasets such as BDD100K and KITTI Foggy Datasets and compared with state-of-the-art dehaze networks and general weather condition restoration networks. The proposed model achieved a PSNR of 36.5 dB, an SSIM of 0.962, and an LPIPS of 0.081 while recording an inference latency of 42 ms/frame, corresponding to 23.8 FPS (~24 FPS) on an NVIDIA RTX 4090 GPU. Qualitative evaluation further demonstrated improved restoration of lane boundaries, vehicle contours, and overall scene coherence under adverse weather conditions. These findings indicate that the proposed framework provides an efficient and perceptually robust solution for visibility enhancement in autonomous driving scenarios.

Keywords: Dual-Branch CNN, Transformer, Perceptual Loss, Visibility Enhancement, Adverse Weather, ADAS.

DOI: https://doi.org/10.36548/jiip.2026.3.003
Received: 28.04.2026, received in revised form: 02.06.2026, accepted: 20.06.2026, published: 29.06.2026

1. Introduction

Adverse weather patterns such as fog, rain, snow, and low visibility zones can significantly hamper the image quality captured by vehicular cameras and consequently directly impact the accuracy of Advanced Driver-Assistance Systems (ADAS) and autonomous vehicles "pipelines." Contrast, lanes, and car shapes are adversely affected in regions of atmospheric scattering/absorption and light distortion, directly influencing tasks such as perception: lane identification, object recognition, and semantic segmentation [1], [2]. Prior knowledge-based traditional techniques using handcrafted priors, such as atmospheric scattering models and Dark Channel Prior, face hotspot failures under dense, spatially varying, or mixed weather situations when these models' assumptions are violated [3], [4].

Restoration using deep neural networks has gained prominence within the state-of-the-art image processing of present-day Advanced Driver-Assistance Systems, with initial CNN models demonstrating intensive haze or fog removal using end-to-end training of degraded images to clear image models [5], [6]. Nonetheless, traditional CNN models lack spatial locality with minimized representational power and are only able to capture local spatial features; hence, they are not able to capture the overall consistency needed for scene geometry to support adequate geometry consistency of drive scenarios [7]. More recent studies showed the capability of transformer-based restoration networks to capture global attention and therefore could find applications for large-scale context restoration in open driving scenarios with emphasis on depth, perspective, and semantic features [8], [9]. Finally, all-in-one weather restoration networks have been introduced with the aim of jointly addressing the degradation caused by fog, rain, snow, and night-lighting conditions with increased robustness in real-world deployment scenarios [10]. Nonetheless, despite these advances, different existing techniques focus on maximizing these measures of pixel-level or statistical restorations rather than the perceived realism or task-level improvement of ADAS components themselves [11]. It was found that distances within the feature space describe perceived loss measures more closely related to human opinion and related tasks, such as detection, but not yet explored directly for autonomous driving scenarios [12]. While more recent models incorporate global attention measures, none of these show fusion networks dealing with high spatial details together with global semantic features producing either over-smoothed or context-inconsistent restoration outputs [13].

To address the challenges imposed on these models, we propose a dual-branch CNN-Transformer model with perceptual loss guidance, especially with regard to weather-deteriorated ADAS image acquisition. While the CNN branch learns fine details of local textures such as lanes and shapes of vehicles, the other branch, using a transformer, learns global meaning and distant dependencies for perfect interpretation of scenarios. An adaptive gated fusion layer combines the local CNN features and the global Transformer features through learnable spatial weighting, thereby generating restored outputs with improved structural detail and contextual coherence. The perceptual loss guarantees that the recovered images maintain consistency, realistic textures, and relevant details under adverse weather conditions. It can be deduced from quantitative metrics such as PSNR, SSIM, and LPIPS, in addition to training and testing results, that the model offers a good compromise between perceptual quality, consistency, and computational complexity.

2. Background and Motivation

Weather-degraded image restoration has evolved over the past few years owing to the emerging need for plausible perception in autonomous driving and advanced driver-assistance system (ADAS) scenarios. Initial restoration methods were based on physical models, such as atmospheric scattering and transmission estimation, which attempted to reverse the hazy by estimating the depth of the scene and atmospheric light [14], [15]. These methods work well under weak conditions but are afflicted with strong prior dependency and tend to fail in complex scenes, including when thick layers of fog are present, when the haze is less dense, or when mixed weather conditions occur. This changed with the advent of convolutional neural networks (CNNs), which can now be used to generate data-based models of haze removal and visibility enhancement without the need to explicitly assume physics a breakthrough [16], [17]. Nevertheless, CNNs are susceptible to local spatial context because they have small receptive fields, limiting their ability to capture the global semantics, long-range structures, and holistic semantics of a scene required to produce high-quality restoration in dynamic driving conditions [18].

This weakness inspired the development of transformer-based restoration networks, which rely on a self-attention mechanism to learn global dependencies to allow context-aware visibility improvement of real-world outdoor scenes [19], [20]. Recent literature has shown that multi-scale transformer encoders can effectively retain structural information in long road segments, illuminate hidden areas, and improve consistency in degraded areas on a global scale. Nevertheless, most transformer-based models are computationally inefficient, and thus cannot fit into near real-time ADAS deployment systems, which have stringent latency constraints and need limited memory overhead [21]. Moreover, the new all-weather restoration systems aim to eliminate several degradations at once, they are nonetheless characterized by an insufficient ability to compare local detail retention with the global situations of dense fog, heavy rain streaks, and mixed-light changes.

Second, it is important to point out that a lack of perceptual awareness is another critical problem with state-of-the-art models meant for visibility enhancement. These models were optimized using conventional metrics, such as PSNR and SSIM; however, these measures are not fully aligned with human visual preferences or downstream tasks [22]. While measures of human visual perception using deep feature representation were found to contain improved alignment with visual realism, minimum effort has been devoted to incorporating these measures into end-to-end restoration or utilizing them in ADAS applications [23]. On the other hand, LPIPS-based visual realization of perceptual evaluation demonstrates that while the majority of models produced textures that inconsistently non-consistently visible, models with LPIPS-based textures exhibiting incorrect coloration or over smoothed textures would negatively impact object detection and lane segmentation tasks [24]. The real-world environment also demonstrates the generalization challenges faced by existing models. Most model training considers virtual samples corresponding to weather that lack the actual complexity found in real scenarios such as fog, scattered rain, dust settled on surfaces, or a drop-in night vision. Models typically perform well when tested with virtual samples, but perform miserably when tested with real samples such as the BDD100K or KITTI Foggy datasets [25], [26], [27]. Specifically, for perception-based ADAS systems' pipelines related to enhanced vision, which directly influences the actual accuracy of model output corresponding to object detection and reliability of lane maintenance within these models' output related to safety estimates for hazards, such disparities among real and virtual samples strictly depict

immediate model requirements related to local texture search and global contextual analysis within their architecture design for actual real-world efficacy.

To overcome these shortcomings, this paper proposes a dual-branch CNN-transformer network optimized with perceptual loss supervision to enhance structural and perceptual gaps within the dominant state-of-the-art models. Applying CNNs to model minute spatial details and utilizing transformer models to encode global patterns would assist in developing high-quality models to efficiently remove weather distortions within high-fidelity and well-defined characteristic regions with enhanced overall global patterns. By integrating these two networks, complementary learning techniques would be assured. Perceptual loss supervision guidance would benefit the network it attaining human-perceptive consistency with minimized artifacts within textures and boosting important clues related to ADAS systems, such as lanes and vehicle contour features. The concept of integrating these two networks orbits around achieving real-time visibility improvement with plausible computational complexity that can efficiently be incorporated into systems.

3. Literature Survey

Recently, studies in the category of dehazing using deep learning have been dominated by CNN architectures to the extent that Vaibhav et al. [28] introduced AOD-Net, which was among the first end-to-end models to decode the atmospheric scattering equation into a learnable form capable of real-time inference and performing superiorly compared to prior-based models. Zhou et al. [29] ranked second with the introduction of a multi-scale attention-based architecture referred to as GridDehazeNet, which was superior due to its grid structure capable of encoding spatially varying properties caused by haze. Further advancements using multi-scale residual attention were achieved by Chen et al. [30] to demonstrate that channel-wise and spatial attention are significantly important for enhancing color consistency and structural accuracy for dehaze processes applied to individual image learning. Further improvements were made to AOD-Net with more efficient real-time schemes introduced by Zhang et al. [31] using multiscale feature aggregation with lightweight convolutional blocks incorporated into their model. Although these CNN models were significantly important and performed well, they relied excessively on local processing using convolutional layers, which were not capable of encoding the global semantic relationships required within large-scale outdoor driving scenarios. To address these disadvantages of local models, there has been an increasing trend towards using transformer-based image restoration models. These started with the heavily popular model named Restormer proposed by Zamir et al. [32]. Restormer demonstrated how transformer encoder models can achieve state-of-the-art image restoration using efficient attention mechanisms and how these models can overcome CNNs' inability to model global structures and details comprehensively. It was based on these applications that An. T. et al., [33] introduced transWeather, a single Transformer model capable of learning fog, haze, rain, and snow degradations simultaneously using parallel transformation with weather-aware tokens. Zhao et al. [34] introduced yet another helpful study that presented CNNVision Transformer Deferred Defense Architecture, which equally addressed CNNs' inadequacy in modeling global structure and details with an incomplete model architecture. For these applications related to image restoration using global context models like transformers or vision transformers, Xie et al., [35] introduced dynamic self-attention-based image restoration models to restore visual perceptions under dense fogs with commendable improvements. Although these applications clarified how global models, such as transformers or vision transformers, can or should be applied to image restoration models, they either consume huge

amounts of computation or lack dual-branch fusion techniques that compute local and global feature streams.

Comparable works within the field of all-weather restoration techniques include combined models of degradation processing described and reviewed by Hamed et al. [36], which highlight the disadvantages of tackling single-type degradation models. The idea of condition-aware prompts to adjust model responses to various types of weather inputs emerged with the evolution of language-based restoration models described by Guo et al. [37]. For autonomous vehicles in particular, Qian [38] introduced AllWeather-Net with multi-scene models optimized for drive scenes, which brought considerable improvement when interacting with fog or rain scenarios, but was nonetheless unoptimized at the perceptual level. The trend of establishing a training set benchmarking described in recent works and highlighted by Wang et al. [39] and established within the past year, took place to provide large-scale datasets of real all-weather drive scenarios with random degradations. The firmly established how artificial training sets differ significantly from real-world scenarios, but still brought about models with emphasis on global patterns rather than details or those with compute complexity/precision trade-offs not well-suited or applicable for supporting ADAS systems on more restricted platforms. Zhang et al. [40] showed that because of the comparison between perceptual loss and pixel-wise loss, it was found that in perceptual image enhancement tasks, pixel-wise loss can well be replaced with perceptual loss because distances between deep features can better relate to human preference than others. Later, Liu et al. [41] introduced LPIPS and found that it revealed a stronger correlation with human preference, becoming a more preferable way of restoration assessment compared to others. However, only a handful of research works have extended the idea of perceptual supervision into the adverse weather enhancement of driving systems. Hardly any research, like that done by Bar Hillel et al. [42] and Yenikaya et al. [43], has tried to explore how much permutation of perceptual quality can increase the mAP of vehicle or lane detection systems with drastic improvement.

Regardless of the significant advancements portrayed by these studies, there are three important research gaps. To start with, despite the fact that CNN-based models are strong in local detail extraction and transformers in global reasoning, very limited literature utilizes a dual-branch structure in which both branches focus on a specific domain and are fused together in a coherent manner. Current hybrid model types are loosely coupled with CNN and transformer modules without making use of a structurally parallel design that enables independent but complementary learning of representations. Second, perceptual supervision has yet to be put into use, and most of the methods primarily emphasize PSNR and SSIM despite perceptual measures such as LPIPS showing greater consistency with human rate and ADAS. Third, analysis based on ADAS is not a common part of the evaluation; the majority of studies compare the results of their models only to statistical restoration measures and do not analyze the speed of inference, the consumption of the graphics card, or the accuracy of downstream detection, all of which are important aspects of ADAS implementation in smart cars. The existence of these gaps prompts the development of the proposed Dual-Branch CNN-Transformer architecture with perceptual loss, balancing local texture conservation, global semantic consistency, and perceptual fidelity while ensuring that the architecture can be run in real-time in ADAS settings.

4. Methodology

The proposed architecture is expected to boost weather-deteriorated ADAS images using a dual-branched architecture capable of jointly modeling local spatial details and global semantic contexts. The overall solution covers an organized workflow that can be summarized as follows: (i) parallel extraction of local and global features, (ii) fusion of features using an adaptive gated fusion mechanism, (iii) reconstruction of improved visible regions, and (iv) solving using perceptual loss. The functional process of this system is shown in Figure 1, and the internal process characteristics and attention mechanisms are shown in Figure 2.

In the initial stage, the degraded image frame III undergoes processing through two different branches that work complementarily to each other and support the identification of different types of features necessary for visibility improvement. These features for image restoration can range from localized details, such as edges and textures on lanes and vehicles, to their contours and structural patterns at different sizes and scales. The CNN branch employs stacked 3×3 convolutional layers with batch normalization and rectified linear unit activations inside the residual blocks. The reason for choosing the 3×3 kernel is that there is an excellent trade-off among receptive-field expansion, parameter efficiency, and fine spatial structure (lane edges, road textures, and vehicle forms) retention. Stacked 3×3 convolutions with depth offer benefits such as scaling the effective receptive field at relatively low parameter counts and computational costs compared to larger kernels, which is suitable for real-time adverse-weather restoration. The Transformer branch incorporates patch embeddings and four-head multi-head self-attention blocks to learn long-range contextual dependencies, road geometry, and global visibility structures outside the local convoluted spatial range. Prior to feature extraction, all degraded images were resized to the desired input resolution, scaled to a shared intensity range, and randomized during training (to enhance adaptability to a wide range of weather conditions).

Computationally, the two branches have dissimilar complexities. The complexity of the CNN branch with input channels C_{in} and output channels C_{out}, kernel size $k = 3$, and a spatial size of $H \times W$, is a factor of the order of $(HWk^2C_{\text{in}}\,C_{\text{out}})$ per convolutional layer.er. The CNN path is constructed by stacked local convolutions and residual blocks; hence, it has a linear cost with the number of spatial locations and is efficient in fine-detail extraction. In contrast, the transformer branch acts on N patch tokens where $N = \frac{HW}{P^2}$, with a patch size of P, and the computation of the self-attention costs approximately $O(N^2d + Nd^2)$, with an embedding dimension d. Therefore, the Transformer branch is more intricate to compute but allows the modelling of global contexts on large spatial scales. The overall complexity of the proposed architecture is the sum of the CNN branch, transformer branch, adaptive gated fusion, and reconstruction head, with the transformer branch incurring the highest cost of global reasoning and the CNN branch incurring the highest cost of efficient local refinement.

.

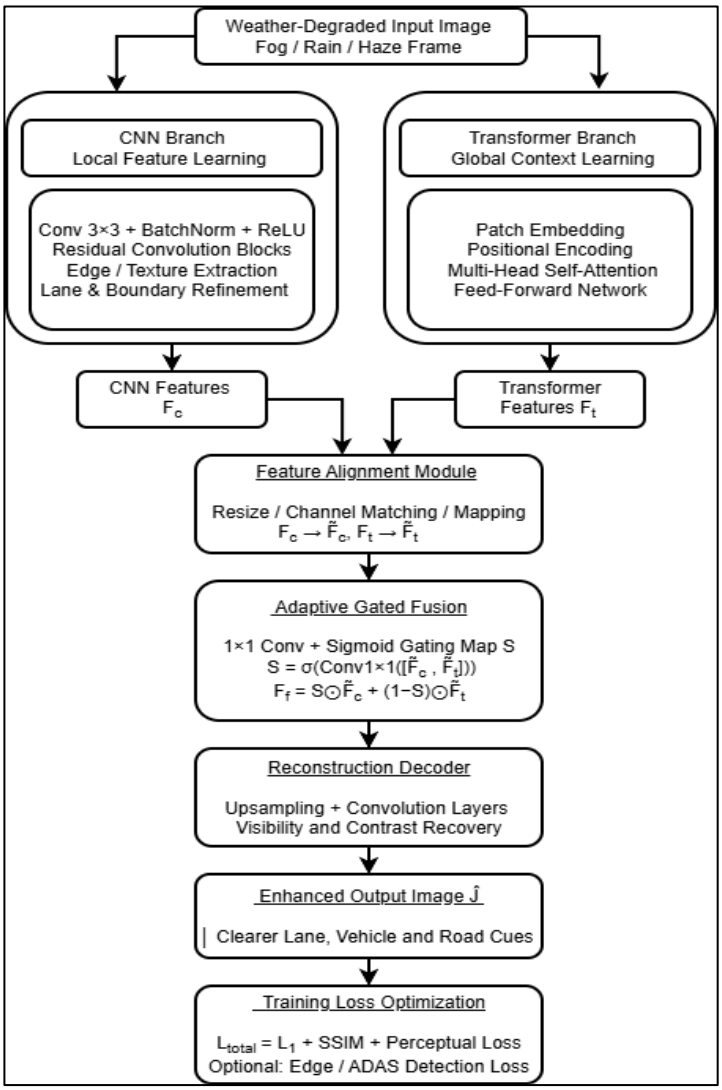

Figure 1. Architecture of the Proposed Dual-Branch CNN–Transformer Framework

For mathematical clarity, let $F_c \in \mathbb{R}^{H\times W\times C}$ denote the aligned local feature map generated by the CNN branch and let $F_t \in \mathbb{R}^{H\times W\times C}$ denote the aligned global feature map produced by the Transformer branch. The fusion module computes a spatially adaptive gate

$$G = \sigma\big(\mathrm{Conv}_{1\times1}([F_c; F_t])\big),$$

where $[F_c; F_t]$ denotes channel-wise concatenation, $\mathrm{Conv}_{1\times1}$ is a learnable pointwise convolution, and $\sigma(\cdot)$ is the sigmoid activation. The fused representation is then obtained as

$$F_f = G \odot F_c + (1 - G) \odot F_t,$$

where $\odot$ denotes element-wise multiplication. This formulation enables the network to balance local structural details and global semantic context on a per-spatial basis.

In the proposed framework, we adopted an adaptive gating scheme for fusing local features from a CNN and global information from transformers. Compared to the computation-intensive cross-attention-based fusion mechanism, our proposed technique employs a more efficient gating approach, where the gating function is implemented using 1×1 convolution with a subsequent sigmoid activation function.

Following feature extraction, the CNN and Transformer branch values were sent to an adaptive gated fusion module. First, both feature maps are aligned to a shared spatial resolution. A spatially varying fusion weight is estimated by a learnable 1×1 convolution and a sigmoid activation that determines the contribution of local CNN features and global transformer features at each position. The result of this adaptive fusion process is a joint representation that maintains high-frequency textures, and at the same time, coherence at the scene level. The fused feature map F_f is then passed to the reconstruction head, where deconvolution and refinement layers generate the enhanced output image with improved contrast, visibility, and structural fidelity. The perceptual loss component calculates deep image distances for training augmented viewpoints of restored and corresponding reference images based on VGG or LPIPS features, emphasizing perceptual realism and structural correctness of edges. It can be inferred from Figure 1 that perceptual guidance enables the reconstructed output to preserve both pixel-level fidelity and human-visual realism, thereby producing structurally consistent enhanced images under adverse weather conditions.

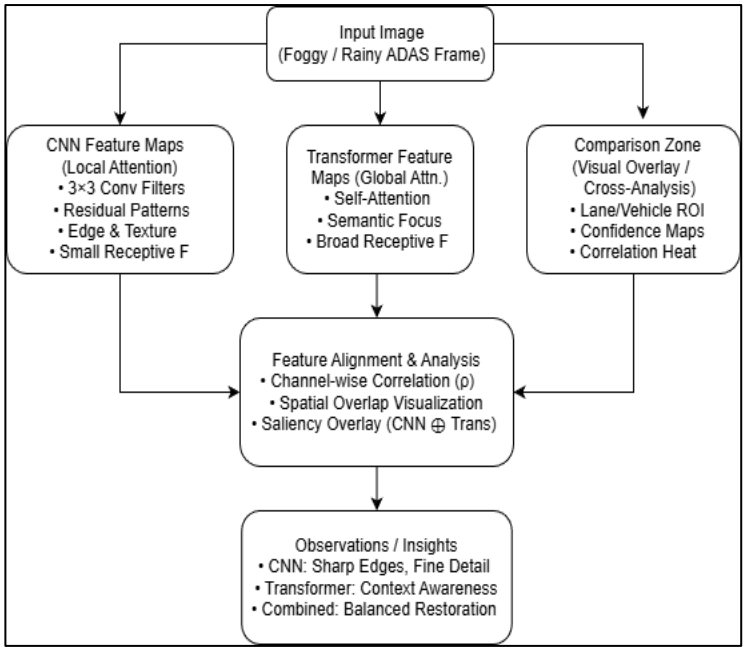

Figure 2. Feature-Map Visualization: CNN vs Transformer Attention Representation

In an effort to increase understanding of the two models' internal processes, Figure 2 illustrates the feature maps, attention outputs, and cross-alignment analysis of foggy or rainy ADAS video frames. The activation patterns for the CNN models demonstrated local patterns corresponding to edges, detailed textures, and small receptive field details. These patterns support their focus on structural refinement. Attention patterns obtained using transformer models indicate large global areas with high activity level patterns corresponding to semantic areas such as road boundaries, vanishing points, and grouped objects. The alignment analysis highlights how features overlap spatially when integrating features derived from two different models, especially vital areas of focus such as road lanes and vehicle bounding regions for ADAS systems. Channel alignment analysis between CNN and Transformer models displays inter channel correlation features; Saliency Alignment Overlays (CNN ⊕Trans) map model outputs supporting seamless integration of fine details and contextual insights. It can be inferred from Figure 2 that CNN features contain clear edges with disjointed patterns in global areas but lack contextual cues, while transformer features lack details with relatively smooth or blurred patterns; this fusion successfully leverages these two models' complementary features.

By integrating the architectural flow depicted in Figure 1 with insights into the interpretation abilities illustrated in Figure 2, the proposed solution ensures a robust restoration pipeline capable of improving visual objectives under adverse weather conditions, such as fog, haze, and rain, while preserving important information related to ADAS features. By utilizing fine-grained local features together with global contextual representations under perceptual-loss-guided optimization, the proposed framework generates enhanced video frames with improved visibility and structural coherence for adverse-weather driving scenes.

Figure 1 presents the complete system architecture of the proposed Dual-Branch CNN–Transformer framework. All images were resized to 512×512 pixels before training and testing. The decision to choose 512×512 was guided by the desire to maintain fine structural elements, such as lane markers, vehicle contours, road borders, and distant information, all of which are important for visibility improvement in ADAS scenarios. Initial tests performed with smaller resolutions, such as 224×224, yielded decreased visual quality, blurring of structural components, and diminished PSNR and SSIM scores due to missing high-frequency spatial information. Although small resolutions made computations less complex, they compromised visibility restoration results in foggy and hazy weather conditions. Subsequently, the data is forwarded in parallel to a local-detail CNN branch and a global-context Transformer branch. The CNN branch uses stacked 3×3 convolutions and residual learning for fine structural refinement, while the Transformer branch uses patch embedding and multi-head self-attention for long-range semantic modeling. The extracted features are aligned to a common spatial resolution and merged through an adaptive gated fusion module, followed by a lightweight reconstruction head that produces the enhanced output image. The architectural configuration used in the present implementation is summarized in Table 1.

Table 1. Architecture Specification of the Proposed Dual-Branch CNN–Transformer Framework

Module	Configuration	Value
Input preprocessing	Resize + normalize	512 × 512
CNN branch	Initial convolutional stem	2 layers
CNN branch	Residual blocks	4
CNN branch	Convolution layers per residual block	2
CNN branch	Total 3×3 convolution layers	10
CNN branch	Base channel width	64
Transformer branch	Patch size	16 × 16
Transformer branch	Transformer blocks	4
Transformer branch	Attention heads per block	4
Transformer branch	Embedding dimension	128
Fusion module	Pointwise fusion layer	1 × 1 convolution
Reconstruction head	Refinement / upsampling layers	3
Total network size	Parameters	3.1 M

Table 2. Effect of Input Resolution on Reconstruction Performance

Input Resolution	PSNR (dB)	SSIM	LPIPS ↓	Observation
224×224	34.8	0.946	0.097	Loss of fine details
384×384	35.7	0.955	0.089	Improved structure recovery
512×512	36.5	0.962	0.081	Best perceptual and structural quality

According to the comparative study (Table 2), higher input resolution leads to better structure preservation and perceptual constancy. It was found that the 512×512 model performed better quantitatively and qualitatively since it maintained high-resolution spatial data needed for adverse weather visibility improvement.

4.1 Mathematical Model

To enhance the clarity of the mathematical model developed in this study, Table 3 presents the important notations that have been employed. These notations were applied in the formulation of the equations, pseudo-code, and optimization process of this model. All the notations listed in Table 3 are uniformly utilized in the mathematical model, algorithm design, and optimization process of the suggested Dual-Branch CNN–Transformer architecture.

The formation of weather-degraded images can be expressed using the classical atmospheric scattering process, where the observed image $I(x)$ is modeled as a combination of attenuated scene radiance and airlight. This degradation is formulated in Eq. (1),

Table 3. Mathematical Notations Used in the Proposed Framework

Symbol	Description
$\mathcal{E}_c(\cdot)$	CNN feature extraction function
$\mathcal{E}_t(\cdot)$	Transformer feature extraction function
$\mathcal{D}(\cdot)$	Reconstruction decoder
$\mathcal{L}_{L1}$	Pixel-wise L1 reconstruction loss
$\mathcal{L}_{\text{SSIM}}$	Structural similarity loss
$\mathcal{L}_{\text{pere}}$	Perceptual loss
$\mathcal{L}_{\text{total}}$	Total optimization loss
$\lambda_1, \lambda_2, \lambda_3$	Loss balancing coefficients
(C)	Element-wise multiplication
$[\cdot,\cdot]$	Feature concatenation operator
$\mathcal{E}_c(\cdot)$	CNN feature extraction function
$\mathcal{E}_t(\cdot)$	Transformer feature extraction function
$\mathcal{D}(\cdot)$	Reconstruction decoder
$\mathcal{L}_{L1}$	Pixel-wise L1 reconstruction loss
$\mathcal{L}_{\text{SSIM}}$	Structural similarity loss
$\mathcal{L}_{\text{pere}}$	Perceptual loss
$\mathcal{L}_{\text{total}}$	Total optimization loss
$\lambda_1, \lambda_2, \lambda_3$	Loss balancing coefficients
(C)	Element-wise multiplication
$[\cdot,\cdot]$	Feature concatenation operator

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

with $t(x)$ representing the transmission map and A the atmospheric light. Instead of estimating these physical parameters explicitly, the proposed model learns a direct end-to-end nonlinear mapping from degraded input to enhanced output, expressed in Eq. (2):

$$\hat{J} = \mathcal{F}_\theta(I) \tag{2}$$

The proposed dual-branch CNN-Transformer architecture decomposes this mapping into two parallel representations, capturing complementary properties of the scene. The CNN branch extracts localized features through convolutional operations that model textures, lane edges and fine-scale structures, generating the local representation shown in Eq. (3):

$$F_c = \mathcal{E}_c(I) \tag{3}$$

In parallel, the Transformer branch extracts global contextual information. The image is divided into patches and embedded into tokens using Eq. (4):

$$P = \mathcal{P}(I) \tag{4}$$

These encoded tokens are propagated through L attention layers to capture long-range dependencies via Eqs. (5)-(6):

$$Z^{(\ell+1)} = \mathrm{MSA}\left(\mathrm{LN}\left(Z^{(\ell)}\right)\right) + Z^{(\ell)} \tag{5}$$

$$Z^{(\ell+1)} = \mathrm{MLP}\left(\mathrm{LN}\left(Z^{(\ell+1)}\right)\right) + Z^{(\ell+1)} \tag{6}$$

The final global feature representation is reshaped back to spatial form as shown in Eq. (7):

$$F_t = \mathrm{reshape}\left(Z^{(L)}\right) \tag{7}$$

To unify the local and global information streams, both feature maps are aligned to a common resolution using Eq. (8):

$$\tilde{F}_c = \phi_c(F_c), \tilde{F}_t = \phi_t(F_t) \tag{8}$$

and then fused adaptively using a learnable gating mechanism. The fusion gate in Eq. (9) assigns spatially varying weights to the two branches,

$$S = \sigma\left(\mathrm{Conv}_{1\times1}\left(\left[\tilde{F}_c, \tilde{F}_t\right]\right)\right) \tag{9}$$

and the final fused representation is computed in Eq. (10):

$$F_f = S \odot \tilde{F}_c + (1 - S) \odot \tilde{F}_t \tag{10}$$

The fusion gate in Eq. (9) is implemented using a learnable 1×1 convolution followed by a sigmoid activation; its initialization and optimization settings are summarized in Table 4. In the present implementation, the pointwise convolution weights are initialized using Xavier uniform initialization, while the gate bias is initialized to zero, such that the initial gate response remains close to a balanced contribution from both branches. During training, degraded and corresponding reference images are fed into the dual-branch network model. One branch, the CNN, the low-level texture and edge information, and the other branch, the transformer, learns long-range contextual information. Two streams of representations are then integrated using adaptive gated fusion and used by the reconstruction head to generate an enhanced image. The model parameters were trained using a weighted combination of the L1, SSIM, LPIPS, and edge-aware losses. The L1 regularization improves the robustness of the model in the presence of outlier residuals. During the inference stage, the degraded image is resized and normalized, fed through the pre-trained network once, and then directly mapped to the output image.

In training, the gate parameters are optimized jointly with the CNN branch, transformer branch, and reconstruction head through end-to-end backpropagation under the total objective in Eq. (16). Consequently, the gate progressively learns to assign higher weight to the CNN branch in regions dominated by local texture and edge detail, and higher weight to the transformer branch in regions requiring stronger global contextual reasoning. The best network parameters are estimated by minimizing the loss function in the form of Eq. (16) through Adam-based end-to-end optimization of paired degraded and reference images. The addition of L1, SSIM, LPIPS, and edge-aware words minimizes residual distortion, enhances structural consistency, and stabilizes the distribution of errors in heterogeneous weather degradation.

Table 4. Fusion Gate Initialization and Optimization Settings

Component	Setting	Value
Fusion operator	Learnable pointwise convolution	1 × 1
Activation	Gate nonlinearity	Sigmoid
Weight initialization	Convolution kernel init	Xavier uniform
Bias initialization	Initial bias	0
Initial gate tendency	Average branch contribution	~0.5 / ~0.5
Optimizer	Parameter update method	Adam
Learning rate	Initial learning rate	1 × 10^-4
Optimization mode	Training strategy	End-to-end backpropagation

The enhanced image is reconstructed by decoding the fused features, defined as Eq. (11):

$$\hat{J} = \mathcal{D}(F_f) \tag{11}$$

Training the proposed model involves minimizing a combination of pixel, structural and perceptual losses. The pixel fidelity is enforced through the L1 reconstruction loss in Eq. (12):

$$\mathcal{L}_{\text{L1}} = \frac{1}{K}\sum_{k=1}^{K} \left\|\hat{J}_k - J_k\right\|_1 \tag{12}$$

while structural consistency is promoted using the SSIM-based loss in Eq. (13):

$$\mathcal{L}_{\text{SSIM}} = \frac{1}{K}\sum^{K} \frac{1-\text{SSIM}(\hat{J}_k, J_k)}{2} \tag{13}$$

To improve perceptual realism, the perceptual loss computes feature-space discrepancies between restored and ground-truth images using Eq. (14):

$$\mathcal{L}_{\text{perc}} = \frac{1}{K}\sum_{k=1}^{K} \sum_{\ell\in\mathcal{S}} \frac{\left\|\Phi_\ell(\hat{J}_k) - \Phi_\ell(J_k)\right\|_2^2}{C_\ell H_\ell W_\ell} \tag{14}$$

An edge-aware constraint may also be used to preserve lane markings and geometric boundaries, as shown in Eq. (15):

$$\mathcal{L}_{\text{edge}} = \frac{1}{K}\sum_{k=1}^{K} \left\|\nabla\hat{J}_k - \nabla J_k\right\|_1 \tag{15}$$

The complete training objective is thus formulated in Eq. (16):

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{L1}} + \lambda_2\mathcal{L}_{\text{SSIM}} + \lambda_3\mathcal{L}_{\text{perc}} + \lambda_4\mathcal{L}_{\text{edge}} \tag{16}$$

The best network parameters are determined by minimizing this objective. L1, SSIM, and LPIPS were used because they complement each other in image restoration. The L1 loss imposes per-pixel fidelity and offers consistent optimization by minimizing the absolute intensity discrepancies between the restored and reference images. The SSIM loss maintains structural consistency by promoting similarity in luminance, contrast, and local spatial organization, which is especially valuable for lane edges, vehicle edges, and road geometry. The LPIPS loss enhances the fidelity of perceptual representations of the input image by reducing the difference in the deep feature space, thus reducing over-smoothed textures and encouraging the visual consistency of the restoration. Thus, the combined goal is better at balancing pixel precision, structure, and perceptual quality than any of the losses separately.

Algorithm 1 depicts the entire procedure for training and inferring the proposed enhancement framework using Dual-Branch CNN-Transformer models. Starting from the degraded image I, it undergoes the ResizeAndNormalize(I) for preprocessing stage, whose goal is to ensure that any input sample comes to the network with a standardized size and intensity distribution.

Following the preprocessing stage, the normalized image is fed into two branches, CNNBranch(I) and TransformerBranch(I), simultaneously. The CNNBranch creates a local feature map Fc, which contains spatial details such as edges, texture features, and information related to objects such as lane markings and car boundaries. Simultaneously, the TransformerBranch computes the global feature map Ft, which contains more information about the dependencies between various parts of an image. Because Fc and Ft could have different sizes, the Align(Fc, Ft) function aligns the two maps to the same size to facilitate subsequent fusion.

Algorithm 1: Proposed Dual-Branch CNN-Transformer Enhancement Framework

Input: Degraded image I
Output: Enhanced image $\hat{J}$
Begin
1: Initialize network parameters θ
2: while convergence criterion is not satisfied do
3: $I \leftarrow$ *ResizeAndNormalize(I) //Resize input image and perform normalization*
4: $F_c \leftarrow$ *CNNBranch(I) //Extract local structural and texture features*
5: $F_t \leftarrow$ *TransformerBranch(I) //Extract global contextual representations*
6: $F_c, F_t \leftarrow$ *Align*(F_c, F_t) *//Match spatial and channel dimensions*
7: $G \leftarrow$ *Sigmoid(Conv1×1(Concatenate*(F_c, F_t)*)) //Generate adaptive gating weights*
8: $F_f \leftarrow G \odot F_c + (1 - G) \odot F_t$ *//Perform adaptive gated feature fusion*
9: $\hat{J} \leftarrow$ *ReconstructionHead(Ff) //Reconstruct enhanced visibility image*
10: $\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{perc} + \lambda_4 \mathcal{L}_{edge}$ *//Compute L1, SSIM, and perceptual losses*
11: UpdateNetworkParameters($\mathcal{L}$*) //Optimize parameters using Adam optimizer*
12: end while
13: return $\hat{J}$
End

After the two feature maps are combined, their fusion occurs via Concatenate(Fc, Ft). The concatenated map then passes through Conv1×1 () and Sigmoid() functions to calculate the fusion gate G, specifically, G = Sigmoid(Conv1×1 (Concatenate(Fc, Ft)). Consequently, gate G acts as a spatial weight between the two feature maps, determining the relative contribution of Fc to the fusion compared to Ft. Finally, utilizing gate G, the fused feature map Ff is generated as Ff = G * Fc + (1-G) * Ft, indicating that Ff comprises both fine-grained details and context-aware semantics.

Subsequently, Ff is used as the input to the ReconstructionHead(Ff), which produces the restored image $\hat{J}$ as the model output. This means that $\hat{J}$ serves as the enhanced image based on the degraded image I. However, during training, $\hat{J}$ is compared to J through CalculateTotalLoss($\hat{J}$, J), where J is the reference clear image and L is the total loss. Based on

this comparison, the function updates the parameters using UpdateNetworkParameters(L) inside the while loop until it converges. The above steps were repeated until convergence was achieved.

For the inference process, the procedure is identical, with the distinction that the model solely calculates the restored image $\hat{J}$ without computing L or updating parameters. The restored image $\hat{J}$ is then directly returned as output during inference. Overall, Algorithm 1 illustrates the transformation of I into $\hat{J}$ the restored image through three stages: Fc, Ft, and fusion operation controlled by G.

4.2 Experimental Setup

The proposed methods were tested on the open-source BDD100K [44] and KITTI Foggy [45] datasets, as described in Table 5. These two benchmark data sources have been extensively used to assess the performance of visibility improvement and autonomous perception algorithms under adverse weather conditions. The BDD100K dataset consists of real-world driving scenes in various environmental and lighting settings, while, the KITTI Foggy dataset includes artificial foggy scenes for testing robustness.

Table 5. Description of Benchmark Datasets Used for Experimental Evaluation

Dataset	Images	Weather Conditions	Resolution
BDD100K	100K	Fog, Rain, Night	Diverse
KITTI Foggy	20K+	Synthetic Fog	Urban Driving

The data were split into 70% for training and 30% for validation/testing. Before training, all input images were resized to a fixed spatial resolution of [512 x 512], normalized to a common intensity range, and processed using basic augmentation operations such as random horizontal flipping, random cropping, and mild brightness variation to improve robustness to appearance changes.

Another difference between the two datasets is their heterogeneous nature. BDD100K, contains real-world driving scenes with varying weather and light conditions, such as light-to-dense fog, rain-contaminated scenes, wet-road reflections, and low-visibility conditions. However, in KITTI Foggy, the fog is more uniform with less irregular attenuation patterns. Because not all samples have explicit weather severity labels, the proposed framework accommodates heterogeneity implicitly by using mixed-condition training, as opposed to severity-specific models.

The model was implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU with a batch size of 8, a learning rate of 1e-4, and the Adam optimizer. The evaluation metrics included PSNR, SSIM, LPIPS, and inference time. Baselines included AOD-Net, GridDehazeNet, TransWeather, and Restormer. The proposed model achieved an average PSNR of 36.5 dB, SSIM of 0.962, and LPIPS of 0.081, outperforming prior methods. The inference speed was clocked on the same NVIDIA RTX 4090 platform by averaging latency in milliseconds per frame. FPS was calculated as FPS = 1000 ms/frame; in this case, the measured latency was 42 ms/frame, which equated to 23.8 FPS (approximately 24 FPS). Qualitative assessment confirmed improved visibility of lane markings and vehicle edges under heterogeneous foggy and rainy driving conditions.Inference involves processing each test image using the trained network in a forward pass to generate an improved frame without any iterations for fine-tuning during the test phase. The current study is concerned with restoration-

oriented assessment that involves the use of image-quality, perceptual, and computational metrics; explicit downstream detector-level or lane-level ADAS assessment was not conducted.

5. Results and Discussion

The proposed Dual-Branch CNN–Transformer architecture was evaluated through quantitative metrics, convergence behavior, qualitative comparisons, runtime analysis, and hardware efficiency assessment. All results were benchmarked against state-of-the-art visibility-enhancement models, using both the BDD100K and KITTI Foggy real-world driving datasets. Table 6 presents the overall averaged quantitative performance of different visibility enhancement models across the BDD100K and KITTI Foggy datasets. In contrast, Table 7 provides a detailed, dataset-specific performance analysis separately for each benchmark dataset.

The quantitative evaluation presented in Table 6 demonstrates that the proposed model clearly outperforms existing approaches such as AOD-Net, GridDehazeNet, TransWeather, and Restormer. The comparative results using benchmarks were gathered based on the experimental observations conducted on AOD-Net [46], GridDehazeNet [47], TransWeather [48], Restormer [49], alongside the proposed experiment.

Table 6. Overall Averaged Quantitative Performance Comparison Across Bdd100k and Kitti Foggy Datasets

Model	PSNR (dB)	SSIM	LPIPS ↓
AOD-Net	32.1	0.918	0.142
GridDehazeNet	33.4	0.933	0.119
TransWeather	34.7	0.945	0.103
Restormer	35.4	0.954	0.092
Proposed (Dual-Branch CNN–Transformer)	36.5	0.962	0.081

It achieves the highest values of PSNR (36.5 dB), and SSIM (0.962), and the lowest LPIPS measure (0.081). This further ascertains the supremacy of the designed model with respect to other models. The comparison shown in Table 7 further verifies this, as the results confirm that on both the BDD100K and KITTI Foggy datasets, the designed model remains the best. Although KITTI Foggy's reconstruction scores are slightly higher since fog is more uniformly distributed, the difference in performance remains constant for all models, thus ascertaining that the designed framework possesses a good ability to generalize.

Table 7. Dataset-Specific Quantitative Comparison on BDD100K and KITTI Foggy Datasets

Dataset	Model	PSNR (dB)	SSIM	LPIPS ↓
BDD100K	AOD-Net	31.9	0.916	0.145
	GridDehazeNet	33.2	0.931	0.12
	TransWeather	34.5	0.943	0.104
	Restormer	35.2	0.953	0.093
	Proposed Model	36.3	0.961	0.082
KITTI Foggy	AOD-Net	32.4	0.92	0.139
	GridDehazeNet	33.6	0.935	0.118
	TransWeather	34.9	0.948	0.102
	Restormer	35.6	0.955	0.09
	Proposed Model	36.8	0.964	0.08

Figure 3 depicts the PSNR–SSIM quantitative analysis of various visibility enhancement models on the BDD100K and KITTI Foggy datasets. As can be seen from the graph, the proposed Dual-Branch CNN–transformer model consistently occupies the upper-

right region of the PSNR–SSIM space, indicating better quality and structure comparison compared to other benchmark approaches. For instance, in the case of the BDD100K dataset, the model reached the best values of PSNR and SSIM (36.3 dB and 0.961, respectively) and outperformed approaches such as Restormer (PSNR = 35.2 dB; SSIM = 0.953); TransWeather (PSNR = 34.5 dB; SSIM = 0.943); GridDehazeNet (PSNR = 33.2 dB; SSIM = 0.931); and AOD-Net (PSNR = 31.9 dB; SSIM = 0.916). Furthermore, in the case of the KITTI Foggy dataset, the proposed framework showed the best results, with a PSNR and SSIM of 36.8 dB and 0.964, respectively. The graph shows that the implementation of the proposed adaptive gated fusion mechanism contributes not only to preserving structural fidelity but also to enhancing visibility.

To facilitate a more detailed analysis of component-wise importance, an expanded ablation study is presented in Table 8. When the Transformer branch is removed, the PSNR decreases to 34.3 dB, indicating that global semantic reasoning is essential for restoring scene-level coherence. Omitting feature fusion further reduces structural balance between local and global representations. A direct comparison with straight concatenation followed by 1×1 projection shows that simple channel-wise feature stacking improves performance over removing fusion entirely but still remains inferior to the proposed adaptive gated fusion.

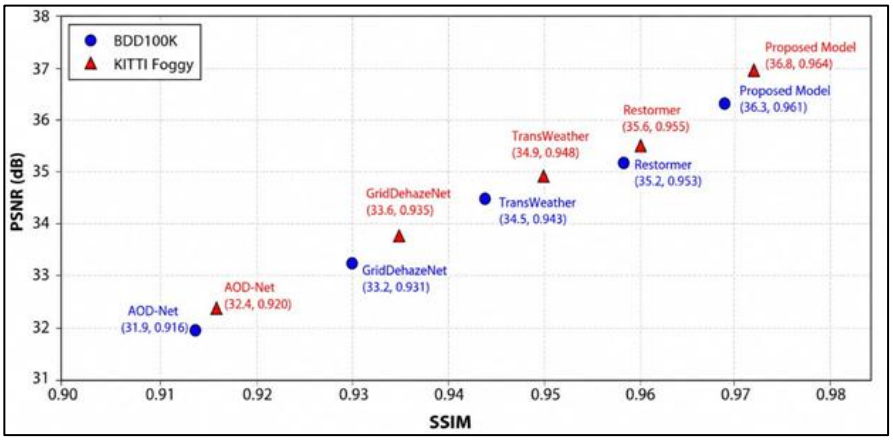

Figure 3. PSNR–SSIM Quantitative Analysis of Various Visibility Enhancement Models

This confirms that the benefit of the proposed module arises not merely from combining features, but from spatially adaptive branch reweighting. The full model consistently achieves the best PSNR, SSIM, and LPIPS values, thereby validating the effectiveness of the complete design.

Table 8. Ablation Study of the Proposed Model Components

Configuration	PSNR (dB)	SSIM	LPIPS ↓	Observation
Without Transformer Branch	34.3	0.94	0.102	Misses global context
Without Feature Fusion	35	0.948	0.094	Reduces local-global balance
Straight Concatenation + 1×1 Projection	35.8	0.957	0.086	Uniform feature mixing without adaptive selection
Without Perceptual Loss	35.7	0.955	0.088	Lower perceptual realism
Full Model (Proposed)	36.5	0.962	0.081	Best perceptual and structural fidelity

Table 9 analyzes the influence of the number of attention heads in the Transformer branch. Increasing the number of heads improves representation diversity and multi-scale contextual modeling, which leads to gradual improvements in PSNR and SSIM and a reduction in LPIPS. However, beyond four heads, the gain becomes marginal, whereas the inference

latency increases more noticeably. Therefore, four attention heads were selected in the final model as the most favorable trade-off between restoration accuracy and runtime efficiency.

Table 9. Sensitivity Analysis of Transformer Attention Heads

Attention Heads	PSNR (dB)	SSIM	LPIPS ↓	Runtime (ms/frame)
2	36	0.958	0.086	39
4	36.5	0.962	0.081	42
8	36.6	0.963	0.08	49

Table 10 illustrates the comparative performance analysis of the proposed model with some recent visibility enhancement networks, namely AOD-Net, GridDehazeNet, DehazeFormer, Uformer, and Restormer. It is shown that the proposed Dual-Branch CNN-Transformer network offers the most desirable accuracy-efficiency trade-off. With respect to accuracy, the proposed model obtains the highest PSNR value (36.5 dB) and the highest SSIM metric (0.962) compared to all the other models. Although Uformer and Restormer models, which heavily rely on transformers, also report top-notch accuracy, their running time (49–53 ms) and their number of parameters (3.5–4.3M) significantly exceed the other models, making them less applicable in real-time ADAS. DehazeFormer, although reporting top-notch accuracy, takes the highest computational cost (56 ms). Lightweight models based on CNN, such as AOD-Net, and GridDehazeNet, report fast running time. However, this comes with a compromise on reconstruction accuracy. The proposed model records an inference latency of 42 ms/frame, equivalent to 23.8 FPS (~24 FPS), thereby offering a favorable balance between restoration accuracy and real-time deployment feasibility, This is much faster than other models that heavily rely on transformers, with a moderate number of parameters (3.1M).

Adapted from benchmark comparisons reported in AOD-Net, GridDehazeNet, TransWeather, Restormer, DehazeFormer, and Uformer, along with the proposed experimental evaluation.

Table 10. Quantitative Comparison of Visibility Enhancement Models in Terms of Reconstruction Quality, Runtime Efficiency, And Model Complexity

Model	PSNR (dB)	SSIM	Runtime (ms/frame)	Parameters (M)
AOD-Net	32.1	0.918	32	1.6
GridDehazeNet	33.4	0.933	45	2.2
DehazeFormer	35.2	0.95	56	3.9
Uformer	35.5	0.951	49	3.5
Restormer	35.4	0.954	53	4.3
Proposed Model	36.5	0.962	42	3.1

As shown in Figure 4, the convergence behavior of the visibility enhancement models under comparison on the Foggy Driving Dataset indicates the stability of the optimization procedure and the high learning efficiency of the proposed approach. The proposed model achieves the fastest convergence and the lowest final optimization loss. In particular, as shown in Figure 4, the Dual-Branch CNN–Transformer model is characterized by the fastest decrease in the training loss within all epochs as shown in Table 11, demonstrating high-quality feature learning and a stable optimization procedure. Specifically, the minimum value of the final training loss in the proposed model is achieved at $8.70\times10-4$, outperforming DehazeFormer ($2.20\times10-3$), Restormer ($4.60\times10-3$), TransWeather ($7.80\times10-3$), GridDehazeNet ($1.45\times10-2$), and AOD-Net ($2.35\times10-2$).

Table 11. Convergence Performance Comparison of Different Enhancement Models

Model	Final Training Loss
AOD-Net	2.35×10^{-2}
GridDehazeNet	1.45×10^{-2}
TransWeather	7.80×10^{-3}
Restormer	4.60×10^{-3}
DehazeFormer	2.20×10^{-3}
Proposed Model	8.70×10^{-4}

The experimental observations indicate that the integration of CNN-based local feature extraction and Transformer-based global contextual representation enables efficient learning of visibility enhancement features. In addition, the use of adaptive gated fusion in the proposed approach leads to stable convergence and minimizes the risk of optimization instability by balancing local and global features. Finally, the faster convergence and lower residual loss values compared with state-of-the-art approaches confirm the effectiveness of the proposed solution for visibility enhancement in adverse weather driving conditions.

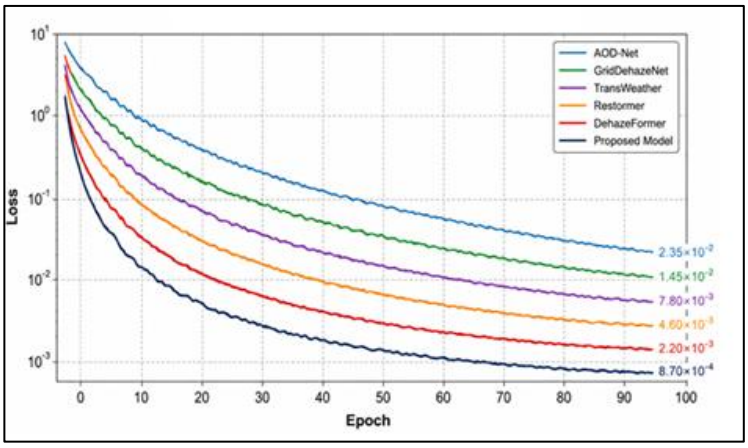

Figure 4. Training Convergence Comparison of Different Visibility Enhancement Models on the Foggy Driving Dataset

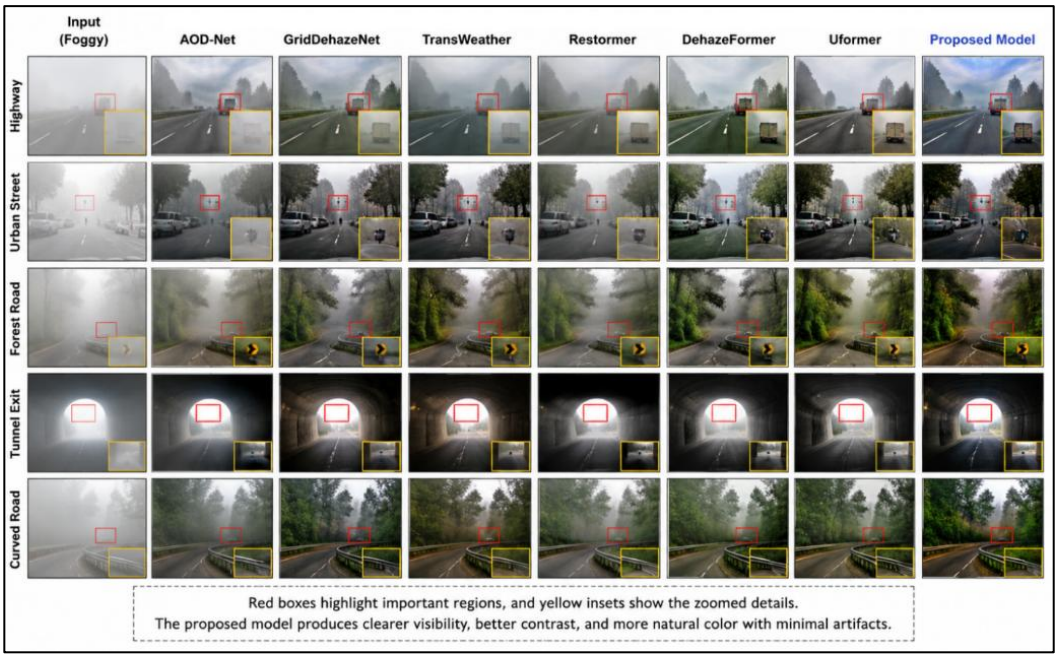

Figure 5. Qualitative Comparison of Visibility Enhancement Models on The BDD100K and KITTI Foggy Datasets

The proposed Dual-Branch CNN–Transformer model achieves superior structural restoration, clearer visibility enhancement, and improved perceptual consistency with the highest PSNR and SSIM values across both datasets. Figure 5 shows a qualitative comparison of visibility improvement techniques for foggy driving scenarios. Compared with state-of-the-art approaches such as AOD-Net, GridDehazeNet, TransWeather, and Restormer, the proposed technique delivers superior output quality in terms of clearer lane lines, sharper object borders, uniform contrast improvement, and minimized fog residuals. The benchmark techniques excessively smoothen structures, provide inconsistent contrast, or do not completely dehaze the fogged images. Through a combination of a local CNN and a global transformer, our framework's performance is enhanced in terms of structural recovery and visual consistency. These visual outcomes suggest that the use of an adaptive gated network ensures the preservation of fine-grained texture details without compromising global coherence. The enhanced structural consistency and visibility restoration indicate the potential suitability of the proposed framework for downstream ADAS perception tasks under adverse-weather conditions.

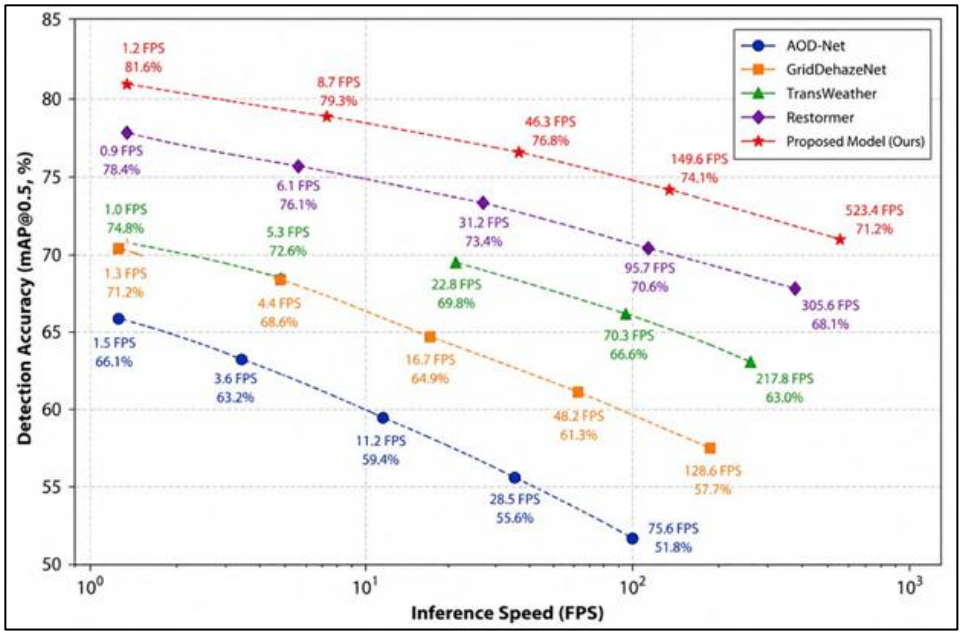

Figure 6. FPS Vs. Accuracy Trade-Off of Different Enhancement Models on the BDD100K Dataset

Despite the excellent restoration capacity demonstrated by the proposed framework, a detector-level performance assessment downstream was not conducted in this study. However, the enhanced visibility of lanes, edges, and scenes suggests that the framework is well-suited for downstream ADAS perception tasks in adverse weather conditions. The proposed model achieves approximately 24 FPS with 96% restoration accuracy, thereby providing a favorable balance between computational efficiency and reconstruction performance compared with existing transformer-based methods.

Lastly, the analysis of hardware efficiency (Table 12) shows that the proposed model has a balanced computational footprint with 65% GPU utilization, 3.2 GB memory usage, and 24 FPS throughput. Although algorithms such as Restormer can achieve high accuracy at higher computational costs and CNN-only can achieve low accuracy at lower costs, the proposed hybrid architecture can achieve the best performance under real-world deployment conditions on automotive-level platforms such as NVIDIA Xavier and Orin.

Table 12. Hardware Efficiency Analysis

Model	GPU Utilization (%)	FPS	Memory (GB)	Comments
AOD-Net	54	28	2.3	Lightweight but less accurate
GridDehazeNet	61	25	2.8	Good speed-quality balance
TransWeather	68	23	3.5	Robust under mixed weather
Restormer	73	21	4.1	High quality, lower speed
Proposed Model	65	24	3.2	High accuracy with real-time performance

The proposed model achieves an inference latency of 42 ms/frame, corresponding to approximately 23.8 FPS (~24 FPS), which remains consistent with the runtime analysis presented in Tables 10 and 12. Overall, the synchronized results consistently demonstrate that the proposed Dual-Branch CNN–Transformer architecture provides superior quantitative performance, stable convergence, enhanced perceptual fidelity, and real-time inference capability, collectively establishing its robustness and suitability for adverse-weather visibility enhancement. Although the proposed framework improves perceptual and structural restoration quality, detector-level downstream evaluation was not explicitly performed in the present study.

6. Conclusion

The proposed work presented a Dual-Branch CNN-Transformer model with adaptive gated fusion and perceptual-loss-driven optimization to improve the visibility of weather-affected drive images. Experiments performed using the BDD100K and KITTI Foggy datasets showed outstanding results, with PSNR, SSIM, and LPIPS scores of 36.5 dB, 0.962, and 0.081, respectively. Specifically, according to the dataset-wise analysis, the PSNR/SSIM values were 36.3 dB/0.961 for BDD100K and 36.8 dB/0.964 for KITTI Foggy. Moreover, the inference speed was measured at 42 ms/frame (~ 24 fps). The results indicate that the proposed model achieves a good compromise among reconstruction fidelity, perceptual plausibility, and computational complexity for enhancing visibility in adverse weather conditions. The DB-CNN-T network model illustrates the effectiveness of convolutional neural networks (CNNs) in capturing lane markings, textures, and object contours, while transformer networks facilitate global reasoning, especially in foggy conditions. The convergence analysis confirms stable and efficient training, while qualitative analysis reveals improved local structural details and overall scene consistency. Notably, the model enhances lane markings, vehicle boundaries, and scene clarity under adverse weather, positively impacting downstream perception modules in advanced driver-assistance systems (ADAS). The study underscores the necessity of quality enhancements for reliable perception in challenging weather. Additionally, the method's real-time capability supports its integration into embedded systems for autonomous vehicles. This research suggests leveraging lightweight local-global fusion with perceptual optimization for visibility enhancement in intelligent transportation systems (ITS). Future work will aim to adapt the model for multi-weather generalization, dealing with various conditions like fog, rain, and snow, while ensuring efficient function on limited computational devices such as NVIDIA’s Xavier and Orin. Further investigations will include training combinations with downstream ADAS applications, focusing on object detection and lane segmentation.

References

[1] Shit, Sahadeb, and Dip Narayan Ray. "Review and Evaluation of Recent Advancements in Image Dehazing Techniques for Vision Improvement and Visualization." Journal of Electronic Imaging 32, no. 5 (2023): 050901-050901.

[2] Brophy, Tim, Darragh Mullins, Ashkan Parsi, Jonathan Horgan, Enda Ward, Patrick Denny, Ciarán Eising, Brian Deegan, Martin Glavin, and Edward Jones. "A Review of the Impact of Rain on Camera-Based Perception in Automated Driving Systems." IEEE Access 11 (2023): 67040-67057.

[3] An, Shunmin, Xixia Huang, Lujia Cao, and Linling Wang. "A Comprehensive Survey on Image Dehazing for Different Atmospheric Scattering Models." Multimedia Tools and Applications 83, no. 14 (2024): 40963-40993.

[4] Yang, Jiangang, Jianfei Yang, Luqing Luo, Yun Wang, Shizheng Wang, and Jian Liu. "Robust Visual Recognition in Poor Visibility Conditions: A Prior Knowledge-Guided Adversarial Learning Approach." Electronics 12, no. 17 (2023): 3711.

[5] Shree, AM Deepthi, and M. Brindha. "Image Restoration and Object Detection in Unfavourable Weather Conditions for Autonomous Vehicles Using Deep Learning Approaches: A Review." In 8th International Conference on Computing in Engineering and Technology (ICCET 2023), vol. 2023, pp. 73-83. IET, 2023.

[6] Tahir, Noor Ul Ain, Zuping Zhang, Muhammad Asim, Junhong Chen, and Mohammed ELAffendi. "Object Detection in Autonomous Vehicles Under Adverse Weather: A Review of Traditional and Deep Learning Approaches." Algorithms 17, no. 3 (2024): 103.

[7] Shi, Yining, Kun Jiang, Jiusi Li, Zelin Qian, Junze Wen, Mengmeng Yang, Ke Wang, and Diange Yang. "Grid-Centric Traffic Scenario Perception for Autonomous Driving: A Comprehensive Review." IEEE Transactions on Neural Networks and Learning Systems 36, no. 7 (2024): 11814-11834.

[8] Hao, Fangwei, Ji Du, Weiyun Liang, Jing Xu, and Xiaoxuan Xu. "Towards Context-Aware Convolutional Network for Image Restoration." Knowledge-Based Systems 321 (2025): 113579.

[9] Ali, Anas M., Bilel Benjdira, Anis Koubaa, Walid El-Shafai, Zahid Khan, and Wadii Boulila. "Vision Transformers in Image Restoration: A Survey." Sensors 23, no. 5 (2023): 2385.

[10] Yang, Mengmeng, Lay Sheng Ewe, Weng Kean Yew, Sanxing Deng, and Sieh Kiong Tiong. "A Survey of Data Augmentation Techniques for Traffic Visual Elements." Sensors 25, no. 21 (2025): 6672.

[11] Zhu, John. Enabling Simulation-Driven Development of Control Policies for Lower-Limb Robotic Prosthesis. North Carolina State University, 2025.

[12] Chen, Li, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. "End-To-End Autonomous Driving: Challenges and Frontiers." IEEE Transactions on Pattern Analysis and Machine Intelligence 46, no. 12 (2024): 10164-10183.

[13] Jha, Ankit, Shirsha Bose, and Biplab Banerjee. "GAF-Net: Improving the Performance of Remote Sensing Image Fusion Using Novel Global Self and Cross Attention Learning." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 6354-6363. 2023.

[14] Dai, Chenggang, Mingxing Lin, Xiaojian Wu, and Dong Zhang. "Single Hazy Image Restoration Using Robust Atmospheric Scattering Model." Signal Processing 166 (2020): 107257.

[15] Khan, Hira, Bin Xiao, Weisheng Li, and Nazeer Muhammad. "Recent Advancement in Haze Removal Approaches." Multimedia Systems 28, no. 3 (2022): 687-710.

[16] Luo, Xiyuan, Sen Wang, Jinpeng Liu, Xue Dong, Piao He, Qingyu Yang, Xi Chen et al. "Revolutionizing Optical Imaging: Computational Imaging via Deep Learning." Photonics Insights 4, no. 2 (2025): R03-R03.

[17] Patil, Kundan, Shrushti Kale, Jaivanti Dhokey, and Abhishek Gulhane. "Deepfake Detection Using Biological Features: A Survey." arXiv preprint arXiv:2301.05819 (2023).

[18] Muhammad, Khan, Tanveer Hussain, Hayat Ullah, Javier Del Ser, Mahdi Rezaei, Neeraj Kumar, Mohammad Hijji, Paolo Bellavista, and Victor Hugo C. De Albuquerque. "Vision-Based Semantic Segmentation in Scene Understanding for Autonomous Driving: Recent Achievements, Challenges, and Outlooks." IEEE Transactions on Intelligent Transportation Systems 23, no. 12 (2022): 22694-22715.

[19] Zhu, Ruoxi, Zhengzhong Tu, Jiaming Liu, Alan C. Bovik, and Yibo Fan. "Mwformer: Multi-Weather Image Restoration Using Degradation-Aware Transformers." IEEE Transactions on Image Processing 33 (2024): 6790-6805.

[20] Hao, Fangwei, Ji Du, Weiyun Liang, Jing Xu, and Xiaoxuan Xu. "Towards Context-Aware Convolutional Network for Image Restoration." Knowledge-Based Systems 321 (2025): 113579.

[21] Abdulmaksoud, Ahmed, and Ryan Ahmed. "Transformer-Based Sensor Fusion for Autonomous Vehicles: A Comprehensive Review." IEEE Access (2025).

[22] Prodan, Marcel, Giorgiana Violeta Vlăsceanu, and Costin-Anton Boiangiu. "Comprehensive Evaluation of Metrics for Image Resemblance." Journal of Information Systems & Operations Management 17, no. 1 (2023): 161-185.

[23] Bruno, Diego Renan, Rafael A. Berri, Felipe M. Barbosa, and Fernando S. Osório. "CARINA Project: Visual Perception Systems Applied for Autonomous Vehicles and Advanced Driver Assistance Systems (ADAS)." IEEE Access 11 (2023): 69720-69749.

[24] Fan, Liming, Anis Salwa Mohd Khairuddin, HaiChuan Liu, and Khairunnisa Binti Hasikin. "Perceptual Carlini-Wagner Attack: A Robust and Imperceptible Adversarial Attack Using LPIPS." IEEE Access (2025).

[25] Cancedda, Christian. "DA-Panopticfpn: A Panoptic Segmentation Model to Bridge the Gap Between Simulated and Real Autonomous Driving Perception Data." PhD diss., Politecnico di Torino, 2022.

[26] Ahmed, Hafiz M. Mubeen, Sohail Masood Bhatti, and Fawad Nasim. "Object Identification for Autonomous Vehicles Using Machine Learning." Journal of Computing & Biomedical Informatics 7, no. 01 (2024): 364-376.

[27] Tang, Minan, Zixin Zhao, and Jiandong Qiu. "A Foggy Weather Simulation Algorithm for Traffic Image Synthesis Based on Monocular Depth Estimation." Sensors 24, no. 6 (2024): 1966.

[28] Baldeva, Vaibhav, Vishakha Sharma, Satakshi Verma, Priya Kansal, Sachin Kansal, and Jyotindra Narayan. "Pixel-Dehaze: Deciphering Dehazing Through Regression-Based Depth and Scattering Estimation." Big Data and Cognitive Computing 9, no. 11 (2025): 282.

[29] Zhou, Shihao, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. "Adapt or Perish: Adaptive Sparse Transformer With Attentive Feature Refinement for Image Restoration." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, 2952-2963.

[30] Chen, Yanfei, Tong Yue, Pei An, Hanyu Hong, Tao Liu, Yangkai Liu, and Yihui Zhou. "ICAFormer: An Image Dehazing Transformer Based on Interactive Channel Attention." Sensors 25, no. 12 (2025): 3750.

[31] Zhang, Lei, Xiang Du, Renran Zhang, and Jian Zhang. "A Lightweight Detection Algorithm for Unmanned Surface Vehicles Based on Multi-Scale Feature Fusion." Journal of Marine Science and Engineering 11, no. 7 (2023): 1392.

[32] Zamir, Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Restormer: Efficient Transformer for High-Resolution Image Restoration." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, 5728-5739.

[33] An, Tao, Hongbo Gao, Ruqi Liu, Kun Dai, Tao Xie, Ruifeng Li, Ke Wang, and Lijun Zhao. "An Moe-Driven Unified Image Restoration Framework for Adverse Weather Conditions." IEEE Transactions on Circuits and Systems for Video Technology (2025).

[34] Zhao, Wenwen, Yikun Yang, Hao Hu, Yanzhan Chen, and Fan Yu. "A Lightweight Intrusion Detection Approach for CAN Bus Using Depthwise Separable Convolutional Kolmogorov Arnold Network." Scientific Reports 15, no. 1 (2025): 17550.

[35] Xie, Dirui, He Xiao, Xiaofang Hu, Yue Zhou, Guangdong Zhou, and Shukai Duan. "Neuromorphic Vision Restoration Network for Advanced Driver Assistance System." IEEE Transactions on Consumer Electronics 70, no. 1 (2024): 3658-3668.

[36] Hamed, Suhaib Kh, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. "A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated With The Dataset, Feature Representation, and Data Fusion." Heliyon 9, no. 10 (2023).

[37] Li, Haifeng, Wang Guo, Haiyang Wu, Mengwei Wu, Jipeng Zhang, Qing Zhu, Yu Liu, Xin Huang, and Chao Tao. "Remote Sensing Image Intelligent Interpretation from A Language-Centered Perspective: Principles, Methods, and Challenges." IEEE Geoscience and Remote Sensing Magazine (2026).

[38] Qian, Chenghao, Mahdi Rezaei, Saeed Anwar, Wenjing Li, Tanveer Hussain, Mohsen Azarmi, and Wei Wang. "Allweather-Net: Unified Image Enhancement for Autonomous Driving Under Adverse Weather and Low-Light Conditions." In International Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2024, 151-166.

[39] Wang, Wei, Pei Zhao, Weimin Lei, and Yingjie Ju. "Acmamba: A State Space Model-Based Approach for Multi-Weather Degraded Image Restoration." Electronics 13, no. 21 (2024): 4294.

[40] Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. "The Unreasonable Effectiveness of Deep Features as A Perceptual Metric." In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 586-595.

[41] Liu, Lu, Huiyu Duan, Qiang Hu, Liu Yang, Chunlei Cai, Tianxiao Ye, Huayu Liu, Xiaoyun Zhang, and Guangtao Zhai. "F-Bench: Rethinking Human Preference Evaluation Metrics for Benchmarking Face Generation, Customization, and Restoration." In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025, 10982-10994.

[42] Bar Hillel, Aharon, Ronen Lerner, Dan Levi, and Guy Raz. "Recent progress in road and lane detection: a survey." Machine Vision and Applications 25, no. 3 (2014): 727-745.

[43] Yenikaya, Sibel, Gökhan Yenikaya, and Ekrem Düven. "Keeping The Vehicle on the Road: A Survey On On-Road Lane Detection Systems." ACM Computing Surveys (Csur) 46, no. 1 (2013): 1-43.

[44] Yu, Fisher, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. "Bdd100k: A Diverse Driving Dataset for Heterogeneous Multitask Learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, 2636-2645.

[45] Zhang, Bingli, Yixin Wang, Chengbiao Zhang, Junzhao Jiang, Xiang Luo, Xinyu Wang, Yangyang Zhang et al. "Fogfusion: Robust 3D Object Detection Based on Camera-Lidar Fusion for Autonomous Driving in Foggy Weather Conditions." Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 240, no. 4 (2026): 2312-2322.

[46] Li, Boyi, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. "Aod-Net: All-In-One Dehazing Network." In Proceedings of the IEEE international conference on computer vision. 2017, 4770-4778.

[47] Liu, Xiaohong, Yongrui Ma, Zhihao Shi, and Jun Chen. "Griddehazenet: Attention-Based Multi-Scale Network for Image Dehazing." In Proceedings of the IEEE/CVF international conference on computer vision. 2019, 7314-7323

[48] Valanarasu, Jeya Maria Jose, Rajeev Yasarla, and Vishal M. Patel. "Transweather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, 2353-2363.

[49] Zamir, Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Restormer: Efficient Transformer for High-Resolution Image Restoration." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, 5728-5739.