

Multi-Label Degradation Classification Of Weather-Affected Images Using Hybrid CNN-Transformer Architecture For Advanced Driving Assistance Systems

Anmol Jain^{1*}, Veerendra Yadav², Harsh Khatter³

^{1,2}*Department of Computer Science and Engineering, Noida International University, Gautam Budh Nagar, Greater Noida, (U.P.), India.*

³*Department of Information Technology, Krishna Institute of Engineering & Technology, Ghaziabad, (U.P.), India.*

**Corresponding author. E-mail: anmolsuvrat@gmail.com*

Contributing authors: vyadav1@ce.iitr.ac.in; harsh.khatter@kiet.edu

Abstract

In the proposed research, the objective of the task is to ensure robust ADAS performance with cameras in adverse weather conditions by being able to determine, for each image, which types of degradation, possibly simultaneously (fog, haze, rain, snow, or their combination), exist. This problem is formalized as multi-label classification with a proposed CNN-transformer network that combines local CNN information with global self-attention capabilities through an image-dependent gate function α . Each individual degradation is represented as an independent Bernoulli distribution with a composite loss function combining binary cross-entropy loss, focal reweighting loss, L2 regularization, and sparse attention, with per-class threshold calibration tailored for ADAS safety preferences. The fundamental contributions of this study are as follows: The weather understanding task for ADAS is modeled as calibrated multi-label recognition instead of single-label scene annotation or simple image restoration. We propose a hybrid ResNet+ViT model featuring an adaptive fusion gate α , which chooses between local and global contexts at the per-image level. We developed a training and calibration pipeline (composite loss function plus class-wise thresholding) to enhance the precision-recall performance, especially for the majority and combined degradations. The study offers a production-ready solution for the front-end, providing degraded-aware indications for routing images to the modules for dehazing, deraining, or desnowing. On the Results On RESIDE, Snow100K, and Dense, testing cross-datasets, we observe that our method outperforms the CNN-only and ViT-only baselines for macro F1 and mAP scores while achieving lower Hamming Loss values and comparable Subset Accuracy/LRAP metric values. The ablation study confirms the effectiveness of α fusion and regularizers, while the FP16/INT8 models stay within the ADAS latency/energy constraints while suffering little accuracy loss. We plan to make our splits and calibrate the thresholds publicly available.

Keywords: ADAS; multi-label classification; adverse weather; CNN-transformer; dehazing; deraining; calibration; attention; RESIDE; Snow100K; DENSE.

1. Introduction

Adverse weather remains a primary failure mode for camera-centric Advanced Driving Assistance Systems (ADAS): fog/haze suppresses scene radiance and contrast, rain introduces high-frequency streaks and veiling, and snow adds granular occlusion—often co-occurring within the same frame. The last few years have established transformers as strong backbones for low-level vision: IPT showed that large-scale pretraining can transfer to imaging tasks [1], while SwinIR demonstrated hierarchical windowed attention for restoration at scale [2]. On the other hand, the NTIRE 2021 Non-Homogeneous Dehazing

Challenge also served as an attempt to assess the range of real-world haze cases, including standard evaluation metrics [3]. To improve the performance in heavy haze, contrastive learning methods have been adopted to improve the compactness and separability of the learned dehazing features [4]. At the same time, Multi-Weather City better considered various weather conditions and domain transfers, which emphasized the point that adverse weather conditions may not exist alone, hence not belonging to one category alone either [5].

As of 2022, significant progress has been made in restoration transformers in terms of their accuracy and efficiency. Restormers are unique because they improve long-range modeling by using an attention mechanism for high-resolution tasks [6]. In dehazing, new models have begun to utilize transmission-aware priors for Vision Transformer (ViT)-based models with a focus on capturing the impact of atmospheric scattering [7]. There are also examples of pure ViT-based baseline models that utilize their global receptive fields for dehazing tasks [8].

On the deraining front, lightweight models such as recursive attention (DRT) and hybrid architectures combining associations with transformers have managed to boost performance without demanding excessive computational resources [9, 10]. The image-de-Raining transformer has even been published in TPAMI, marking it as mature and sufficiently reliable for practical deployment [11]. Other innovations, such as multiscale or channel-specific transformers and pyramidal attention mechanisms, further enhance robustness against various types of rain [12, 13].

Improvements in the quality of supervision also had an impact; new ground-truth deraining protocols highlighted the limitations of working only with synthetic examples [14]. Putting all the pieces together, the TransWeather model shows that a transformer can handle multiple harsh weather conditions in a single framework and that various degradations might share a common global structure [15].

Although recent progress has been made, many existing methods still rely on the assumption of a single dominating degradation and are optimized either for the image quality measure PSNR or SSIM or for the accuracy of a single label classification task. However, tasks such as ADAS require more subtle skills such as:

1. Multilabel recognition of fog, haze, rain, snow, and their mixtures.
2. An adaptation to achieve a balance between local details (for example, rain streaks or snow pixels) and the overall context (for example, a haze veil).
3. Per-class thresholds to facilitate decision making with a focus on safety.

These requirements drive the construction of a hybrid CNN and transformer architecture, which incorporates the best qualities of each model. The architecture incorporates a mechanism for the fusion of local and global information learned, producing Bernoulli probabilities for each weather class independently. Real-world conformity was incorporated into the model via a class-wise threshold-calibration component. The primary key contributions of this study are described below.

- Hybrid CNN and transformer model for multi-label weather classification in ADAS, where adaptive fusion gate combines local and global knowledge according to image content.
- The output layer has a probabilistic structure consisting of independent Bernoulli heads, and composite loss is employed to handle the problem of class imbalance and attention regularization.
- A procedure for evaluation that extends current restoration evaluation practices by considering the quality of model calibration and robustness, which is more aligned with model requirements in a safety-critical setting.

2. Background and Motivation

Between 2023 and 2024, considerable progress was made in the restoration and recognition of adverse weather conditions, thanks mostly to transformer-based models. These models improve the ability to grasp the long-range context without losing local details. In the de-raining process, sparse and structured attention techniques simultaneously improve output quality and computational complexity [16]. A unified model for handling raindrops and rain streaks has been proposed, which adopts an uncertainty-guided ranking scheme to emphasize degraded areas in the process [17].

For dehazing, multibranch transformer architectures based on Taylor series expansion were proposed for better capture of the scattering patterns on varying spatial scales [18]. Simultaneously, efficient models, such as the dilated convolutional transformer, have also been proposed to provide high performance without a high computational cost [19]. Strategies involving the alternative use of attention mechanisms in spatial or channel domains have helped improve the adaptability of the methods to varying patterns for heavy rain conditions [20]. Nonetheless, the results of benchmarking studies have confirmed the presence of domain gaps, which have emphasized the challenges faced by methods in generalizing to real-world data environments beyond the synthetic data settings represented by current puddle synthesis methods [21].

For weather categorization, Vision Transformer (ViT) backbones were shown to be competent, albeit comparatively restricted to single-label outputs and very broad categorization, thereby undermining their general applicability to real-world driving tasks, which involve a combination of weather patterns [22]. However, more general restoration models have been proposed for dealing with weather-related corruptions, including residual dense grid transformers, which provide an efficiently parametrized solution [23]. On the other hand, more specialized advancements, such as de-snowing models with dedicated backbones and learning by prototype approaches, have made de-snowing more resilient to finer visual obstructions [24-26].

Two comprehensive real-world dehazing efforts demonstrated that non-homogeneous haze, night-time scenes, and device variability remain challenging, even for strong transformers [27, 28]. Community evaluations (NTIRE 2024) codified dense/non-homogeneous haze protocols and metrics, underscoring the need for rigorous, field-relevant testing [29]. Finally, ADAS-oriented work integrating weather cues with detection confirmed that degradation awareness benefits downstream perception but still relies on simplified labels and heuristics rather than calibrated multi-label decisions [30].

The literature converges on three gaps relevant to ADAS: (i) most systems are single-degradation restorers or single-label classifiers unsuited to co-occurrence; (ii) fusion of local textures (streaks/flakes) and global veiling (haze/fog) is often fixed rather than adaptive per scene; and (iii) evaluation/reporting typically emphasizes PSNR/SSIM or coarse accuracy, not calibrated, per-class operating points required for safety policies. These observations motivate our hybrid CNN transformer with adaptive fusion and multi-label Bernoulli outputs, paired with class-wise threshold calibration to expose actionable, degradation-aware decisions for downstream ADAS modules.

3. Literature Survey

Recent work is most relevant to multi-label weather understanding for ADAS clusters into four themes: (i) transformer dehazing with physics/channel priors, (ii) deraining with multi-scale prompting and semi-supervision, (iii) hybrid CNN-transformer backbones for multi-label recognition, and (iv) ADAS-oriented robustness.

Recent advancements have continued to strengthen the case of hybrid models that integrate both local and global information in adverse-weather recognition and restoration tasks [31]. For instance, MCRFS-Net introduces multi-scale contrastive regularization and frequency selection techniques that improve the separation of low-frequency haze veils from high-frequency scene details under non-uniform haze conditions [32]. Complementing this, a deep transfer strategy combined with local mean adaptation enhances the color fidelity and stability in real-world scenarios with spatially varying haze [33]. A codebook prior-guided hybrid attention mechanism has been shown to better regulate long-range dependencies and minimize artifacts, resulting in improved color constancy across diverse scenes [34]. Together, this works make a compelling case for a representation that fuses global features (as modeled by ViTs) with local cues (captured by CNNs), exactly the type of synergy that our hybrid architecture aims to achieve.

Within the de-raining context, the MSPformer incorporates a conditioned prompt to dynamically adjust attention based on rain type (fine streak rain and heavy rain) while ensuring efficiency. This has shown robust performance on conventional benchmarks [35]. In environments where the target images are not available, a semi-supervised framework involving wavelet Swin is employed to leverage the rainy images and maintain the level of performance for the restoration task [36]. This continues to emphasize the need to ensure data efficiency in the learning objectives for weather-aware recognition systems.

More evidence for the hybrid architecture comes from multilabel classification studies outside the area of autonomous driving. The CNN-transformer hybrid model with multilabel learning achieved better results when processing sensor inputs and dealing with simultaneous conditions compared to the conventional single-stream architecture [37]. This reinforces the correctness of the proposed idea to leverage the independent Bernoulli components and the adaptive fusion gate to identify simultaneous conditions such as fog, haze, rain, and snow.

It is often the metrics related to peak restoration that attract attention in the literature. However, any actual Advanced Driver Assistance System requires more than merely image quality. It must be resilient across a domain change and capable of seamless integration with any on-board sensing solution. A recent work in the IEEE Open Journal of Vehicular

Technology demonstrated that a combination of deep models with a thermographic sensor is very promising for perception under adverse weather conditions [38].

Citation	Task / Theme	Problem Framing	Method / Core Idea	Data / Setting	Key Findings	Gap
Wu et al., [31], 2025	Derainin g	Rain removal with varied streak density	Multi-scale Transformer with conditioned prompts (MSPformer) to adapt attention by rain type	Standard derain benchmarks	Robust, efficient deraining across fine/heavy rain	Restoration-only; no multi-label recognition or calibrated thresholds
Hao et al., [32], 2025	Semi-supervised derainin g	Limited labels / no pairs	Wavelet decomposition + Swin Transformer; semi-supervised on rainy images	Real/synthetic rainy, unpaired	Competitive quality under label scarcity	Not frame-wise multi-label; lacks ADAS operating-point control
Chai et al., [33], 2025	Hybrid derainin g	Non-uniform weather artifacts	Hybrid CNN-Transformer attentive network fusing local detail with global context	SPIE real-world scenes	Better color fidelity/stability under non-uniform conditions	No multi-label head; no per-class calibration
Li et al., [34], 2025	Dehazin g	Physics-consistent haze removal	Physics-based image-formation priors + Transformer to regularize long-range attention	Adverse haze scenes	Improved structure preservation, fewer artifacts	Restoration metrics focus; no safety-aligned decision thresholds
Zou et al., [35], 2025	Derainin g rep. learning	Cross-paradigm generalization	Cross-Paradigm Representation & Alignment	Public derain sets	Strong cross-benchmark generalization	Lacks multi-label outputs; no threshold

			Transformer to align features across degradations			calibration
Vasilescu et al., [36], 2025	Dehazing robustness (i/iv)	Robustness to distribution shifts	Adversarial-robust ViT fine-tuning for stability under perturbations	Adverse-weather dehazing	Improved robustness vs. standard ViT	Restoration-centric; not a calibrated recognition front-end
Nawaz et al., [37], 2025	ADAS perception (iv)	Visibility loss in extreme weather	Deep RGB+thermal fusion for weather-resistant perception	Vehicular datasets	Promising ADAS gains under harsh weather	Requires thermal HW; our method is camera-only recognition with calibration
Huang et al., [38], 2025	Dehazing w/ priors (i)	Artifact control / color constancy	Codebook prior-guided hybrid attention to regularize long-range dependencies	Diverse outdoor scenes	Improved color constancy, fewer artifacts	No multi-label classification; lacks per-class thresholds

In response to these practical demands, our evaluation framework incorporates class-wise thresholding and cross-dataset validation to better reflect real-world deployment scenarios.

Collectively, the existing literature motivates the following design principles:

1. Fusion of global attention and local detail for the effective handling of haze and rain
2. Prompted or semi-supervised learning to address label scarcity under diverse weather conditions
3. Hybrid CNN-Transformer architectures to support multi-label weather classification;
4. An overarching focus has been placed on the ADAS-aligned robustness and interpretability.

Our method combines these principles through a unified architecture that includes the following:

- An adaptive fusion gate to combine local and global features dynamically,
- Bernoulli output heads paired with a composite loss function (combining likelihood, focal loss, and regularization),
- In addition, per-class calibration mechanisms are used to align the predictions with the safety requirements of real-world ADAS applications.

4. Proposed Methodology

We propose a hybrid CNN–Transformer framework, which pairs global, context-aware representations and local, texture-related features, for separating a number of simultaneous co-occurring driving image degradations resulting from poor weather conditions. End-to-end design (pipeline, network, and training-calibration) is summarized in Figures 1–3, explaining the system overview, dual-path structure and adaptive fusion, and optimization and post-hoc calibration workflow, correspondingly.

4.1. System overview of the proposed multi-label weather recognition front-end for ADAS

A camera frame is preprocessed and encoded in **two** parallel pathways: a local CNN Φ^{cnn} that captures high-frequency textures (e.g., rain streaks, snow granules, lane edges) and a global transformer T that captures long-range context and veil-like structures (e.g., fog, haze, illumination). An adaptive gate α blends both streams into a fused representation z , which drives multi-label heads to score fog, haze, rain, snow, and mixed conditions. Class-wise calibration selects operating thresholds $\tau(c)$ to yield binary labels that trigger a restoration router (dehaze/derain/desnow) and feed downstream ADAS consumers (detection, segmentation, and planning). A composite training objective shapes the encoders and gate, and a runtime confidence score κ enables a conservative fallback when predictions are uncertain.

Figure 1 summarizes the end-to-end pipeline that converts a single camera frame into actionable Multi-label weather decisions for driving. After standard preprocessing, the image was processed using two complementary encoders. The local pathway Φ^{cnn} emphasizes high-frequency cues that are diagnostic of precipitation and fine structural detail, while the global pathway T models spatially extensive context and low-frequency attenuation characteristics of fog and haze. The outputs are fused by an image-conditioned gate α that assigns channel-wise weights to the local and global features, allowing the system to adapt per frame, for instance, favoring global context in heavy haze and local textures during dense rain or snow.

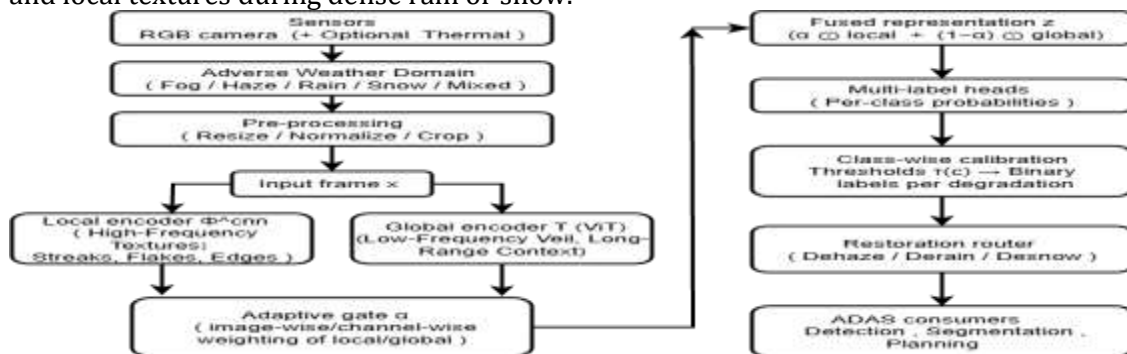


Figure 1. System overview of the proposed multi-label weather recognition front-end for ADAS

The fused representation z is passed to multi-label heads that produce independent probabilities for each degradation, accommodating co-occurring conditions within the same frame. To obtain deployment-ready decisions, the system applies class-wise calibration to determine thresholds $\tau(c)$ on a held-out set, translating probabilities into binary labels with the desired precision-recall trade-off. These labels inform a restoration router that activates only the necessary enhancer (dehaze/derain/desnow), preserving latency and computing for downstream tasks. The cleaned or flagged outputs are then consumed by the ADAS perception and planning components.

Model behavior is shaped during training by a composite objective that balances likelihood fitting, class-imbalance handling, and regularization of attention/distributed focus. At inference, a scalar confidence κ summarizes the margin of the predictions across classes; when κ is low, the system can abstain, aggregate temporally, or route to a restoration-first path, supporting a safety-critical operation. Together, the dual-path encoding, adaptive fusion, and calibrated multi-label decision rule explain the method's improvements in fog, haze, rain, snow, and mixed conditions across the datasets and operating domains.

4.2. Hybrid CNN-Transformer architecture for multi-label weather recognition

The input image x is processed as shown in figure 2 by two complementary encoders: a local CNN pathway Φ^{cnn} that emphasizes high-frequency textures (e.g., rain streaks, snow granules, lane edges) and a global transformer pathway T that models long-range spatial context and low-frequency attenuation (e.g., haze and fog veils). The ViT branch receives tokens obtained via patchification and linear projection with positional encodings and is composed of repeated MHSA + FFN blocks with residual connections and normalization. Each branch yields a compact feature vector f^{cnn} and f^{vit} , which are fused by an adaptive gate α learned from their joint evidence. The fused representation z drives multi label heads to produce independent probabilities for fog, haze, rain, snow, and mixed conditions.

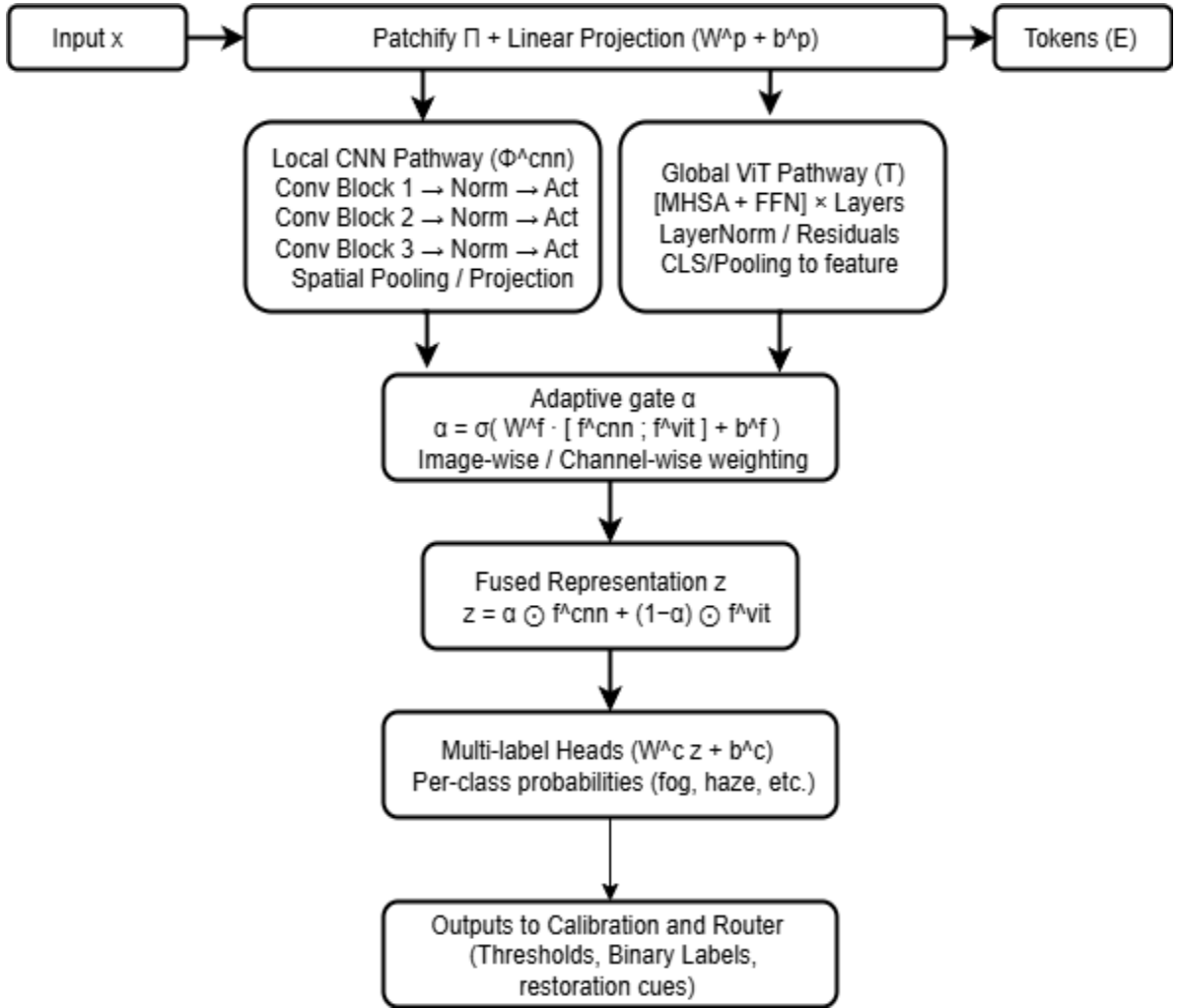


Figure 2. Hybrid CNN-Transformer architecture for multi-label weather recognition

Image x is routed to two pathways. The local pathway Φ^{cnn} follows a ResNet-style stack of convolution–normalization–activation blocks with residual connections; early strides capture fine textures while preserving sufficient spatial detail. The final pooling/projection stage yields a vector $f^{cnn} \in R^D$. In parallel, the global pathway T first patchifies x into non-overlapping $p \times p$ patches and linearly projects each patch to a token embedding; positional encodings are added to retain the spatial order. The token sequence is processed by L layers of multihead self-attention and feed-forward networks (pre-norm with residuals). A CLS token or mean pooling produces $f^{vit} \in R^D$. This design separates high-frequency evidence (precipitation, edges, signage microstructure) from low-frequency context (contrast/visibility gradients and large-scale illumination), matching the multi-scale nature of adverse weather.

To combine the streams, the model learns an image-conditioned gate $\alpha \in [0,1]^D$ from the joint evidence of f^{cnn} and f^{vit} using a light MLP. The gate operates channel-wise, enabling per-dimension emphasis on local or global cues. This convex, continuous fusion

avoids brittle hard switches and allows the network to lean into a global structure during heavy haze while prioritizing local textures when rain or snow dominates.

The fused representation $z \in R^D$ feeds independent linear heads with sigmoid activations, producing a probability for each target degradation (fog, haze, rain, snow, and mixed). Because degradations may co-occur in a single frame, the heads are not coupled by softmax. These probabilities were later calibrated and threshold (see Fig. 3) to yield deployment-ready labels for restoration routing and downstream ADAS perception.

In practice, we match the feature dimension d across branches (e.g., $d=512-768$) to simplify fusion; typical choices are patch size $p=16$ and depth $L=8-12$ for the transformer, with global average pooling on the CNN side. Lightweight choices for the gate (one or two hidden layers) keep overhead minimal. Normalization (Batch/Layer Norm) and dropout were applied in their standard places. This configuration yielded strong accuracy–efficiency trade-offs in our ablations, with the gate and dual-path designs jointly responsible for consistent gains under fog, haze, rain, snow, and mixed conditions.

4.3. Training objective and class-wise calibration workflow

Model outputs are supervised by a composite objective that combines per-label Bernoulli likelihood with a focal term to emphasize rare cases and regularization to stabilize weights and sharpen attention. Gradients update both encoders and the fusion gate. After each epoch, a held-out set is used to calibrate class-wise operating thresholds that convert probabilities into binary decisions; these calibrated thresholds are stored for deployment and paired with a runtime confidence score to support conservative fallbacks in safety-critical settings.

For each image and label, the model as shown in figure 3 predicts an independent probability; losses are accumulated in a multi-label regime (no softmax coupling). The objective contains (i) a per-class Bernoulli likelihood that anchors probabilities to the ground truth; (ii) a focal term that down-weights easy positives/negatives and focuses learning on minority labels and ambiguous cases; and (iii) lightweight regularization—weight decay for parameter stability and an attention-sparsity prior that discourages diffuse, non-diagnostic focus. This combination proved effective for long-tailed label distributions and co-occurring degradation.

We train end-to-end (CNN branch, transformer branch, and fusion gate) with Adam or SGD, gradient clipping to prevent exploding updates, and a cosine schedule or step decay for the learning rate. Mixed-precision is used where available to satisfy the memory and throughput constraints. Early stopping monitors validation macro-F1 to avoid overfitting while respecting the class balance of the application.

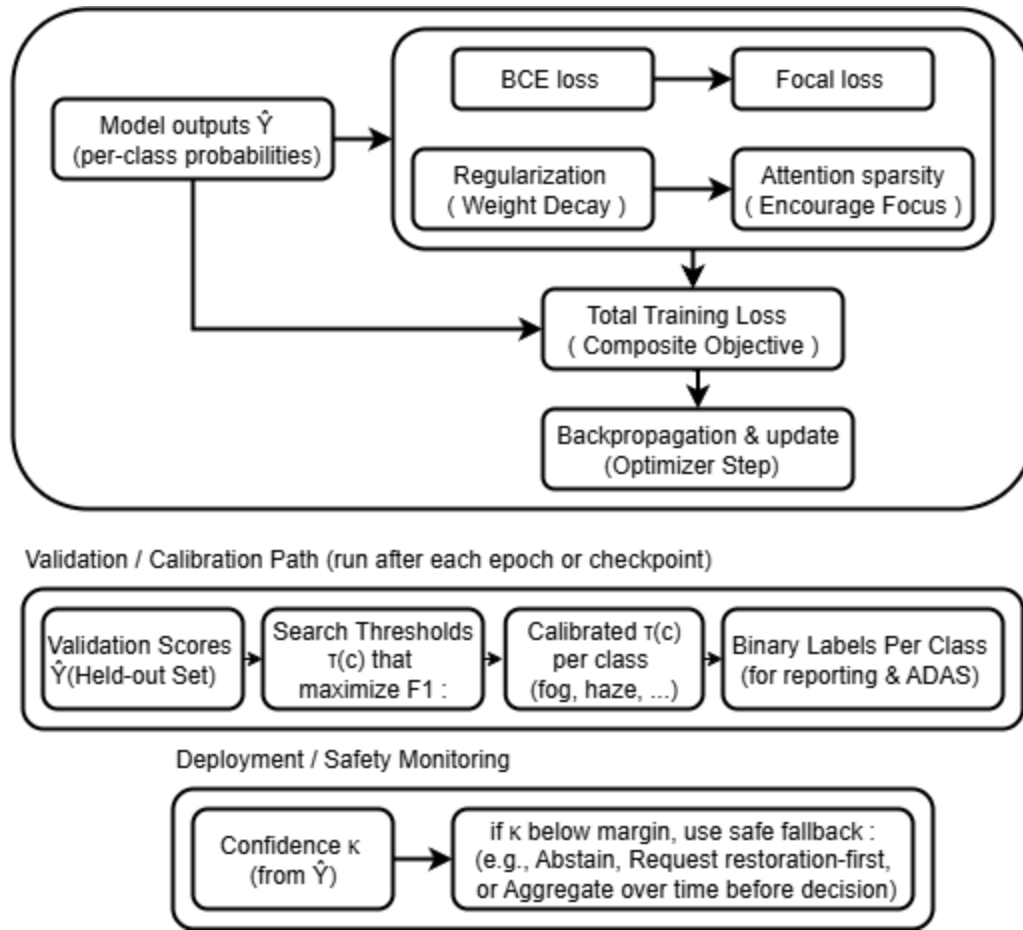


FIGURE 3 Training Objectives and Calibration Workflow

Because multi-label probabilities are not inherently calibrated, we tune a per-class threshold $\tau(c)$ on a held-out validation set after each epoch (or at the training end). For each class, a coarse-to-fine search over the unit interval selects the threshold that maximizes validation F1 (other operating points, precision-favored or recall-favored, can be chosen if an application demands). The resulting set $\tau(c)$ persists with the checkpoint and is used at inference to map probabilities to binary labels. This decouples representation learning from operating-point selection, yielding stable precision–recall trade-offs across datasets.

At runtime, the classifier emits both per-class probabilities and binary decisions obtained through $\tau(c)$. A scalar confidence summary κ (e.g., the average distance of probabilities from 0.5) supports conservative behavior. If κ falls below a preset margin, the system can abstain, aggregate temporally, or route frames through a restoration-first path before downstream use. In practice, this calibration-plus-confidence scheme reduces false co-activations in mixed weather and aligns the operating point of the detector with ADAS risk constraints.

The composite objective drives discriminative features across heterogeneous weather scales (local streaks/granules vs. global veils), whereas calibration ensures that the decision rule reflects application priorities rather than raw score distributions. Together, they

deliver the observed gains in macro-F1 and mAP, lower Hamming loss through fewer spurious multi-labels, and predictable behavior under domain shift

4.4. Mathematical Model

Let $x_i \in \mathbb{R}^{H \times W \times 3}$ denote an RGB frame and $y_i \in \{0,1\}^C$ its multi-label indicator over $C = 5$ degradations { fog, haze, rain, snow, mixed }. We learn two representation functions- Φ^{cnn} for local textures and T for global context-and a data driven fusion that yields a single embedding z_i used by independent Bernoulli heads.

$$f_i^{cnn} = \Phi^{cnn}(x_i; \theta^{cnn}) \quad (1).$$

The convolutional encoder maps x_i to a local, high-frequency representation f_i^{cnn} as in Eq. 1 governed by parameters θ^{cnn} . Its finite receptive fields sharpen cues such as streaks (rain) and granules (snow), which later complement global context.

$$E_i = W^p \Pi(x_i) + b^p \quad (2)$$

The patchify-project step as in Eq. 2 produces a token sequence E_i suitable for self-attention. Operator $\Pi(\cdot)$ slices x_i into non-overlapping patches; W^p, b^p lift each patch into a latent space that preserves coarse spatial layout correlated with haze/fog.

$$\begin{aligned} Q &= EW^q, K = EW^k, V = EW^v, \text{Attn}(Q, K, V) \\ &= \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \end{aligned} \quad (3)$$

Scaled dot-product attention as shown in Eq. 3 forms long-range dependencies by comparing all tokens. Queries select content via similarity to keys; normalized weights aggregate values into globally informed features, which are crucial for diffuse, low frequency degradations (fog/haze).

$$f_i^{vit} = T(E_i; \theta^{vit}), A^{\ell, h} \\ = \text{Attn}^{\ell, h} \quad (4)$$

The transformer T stacks multi-head attention and pointwise feedforward layers to produce a global representation f_i^{vit} as shown in Eq. 4. Per-layer, per-head matrices $A^{\ell, h}$ are retained to regularize attention (sparsity) and curb spurious, diffuse focus.

$$\alpha_i = \sigma(W^f [f_i^{cnn}; f_i^{vit}] + b^f) \quad (5)$$

In Eq. 5 the adaptive gate $\alpha_i \in [0,1]^d$ is a learned, image-conditioned weight (optionally channel-wise) that decides the local-global balance. Large entries tilt toward local textures; small entries favor global context. This replaces fixed fusion with a scene-aware convex blend.

$$z_i = \alpha_i \odot f_i^{cnn} + (1 - \alpha_i) \odot f_i^{vit} \quad (6)$$

The fused embedding z_i is the elementwise convex combination of local and global features as per Eq. 6. Because $\alpha_i \in [0,1]^d$, each channel of z_i lies in the convex hull of the corresponding channels of (f_i^{cnn}, f_i^{vit}) , ensuring stable interpolation rather than brittle switching-useful for co-occurring weather.

$$s_i = W^c z_i + b^c, \hat{y}_i(c) = \sigma(s_i(c)) \quad (7)$$

Linear heads map z_i to per-class logits s_i and probabilities $\hat{y}_i(c) \in (0,1)$ as per Eq. 7. Using sigmoids instead of softmax abandons mutual exclusivity and models each class as an independent Bernoulli, which is appropriate when degradations co-occur in a single frame.

$$L^{bce} = \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C [-y_i(c) \log \hat{y}_i(c) - (1 - y_i(c)) \log (1 - \hat{y}_i(c))] \quad (8)$$

planation (Eq. 8). The binary cross-entropy maximizes the per-class Bernoulli likelihood across images and labels. It is the proper scoring rule for independent labels and serves as a baseline term before imbalance and structural regularizers are introduced.

$$L^{foc} = \frac{1}{NC} \sum_{i,c} \left(-\alpha^{foc} (1 - \hat{y}_i(c))^\gamma y_i(c) \log \hat{y}_i(c) - (1 - \alpha^{foc}) \hat{y}_i(c)^\gamma (1 - y_i(c)) \log (1 - \hat{y}_i(c)) \right) \quad (9)$$

The focal term as per Eq. 9 reweights the BCE so that easy cases contribute less and hard, under-represented cases contribute more. Exponent γ controls the strength of focusing; α^{foc} balances positive/negative prevalence, improving macro-averaged metrics when classes such as mixed are rare.

$$L^2 = \|\theta^{cnn}\|_2^2 + \|\theta^{vit}\|_2^2 + \|W^f\|_2^2 + \|W^c\|_2^2, L^{att} = \frac{1}{LH} \sum_{\ell,h} \|A^{\ell,h}\|_1 \quad (10)$$

Explanation (Eq. 10). The ℓ_2 penalty contracts parameter magnitude to reduce overfitting, while the ℓ_1 penalty on attention encourages sparse, focused heads rather than diffuse patterns. Together they temper capacity and guide the allocation of attention to salient regions as shown in Eq. 10.

$$L^{tot} = L^{bce} + \lambda_1 L^{foc} + \lambda_2 L^2 + \lambda_3 L^{att} \quad (11)$$

The composite objective as shown in Eq. 11 trades off likelihood fit, imbalance robustness, and structural regularization. Hyper-parameters $\lambda_{1..3}$ are selected on validation to optimize deployment-relevant criteria (e.g., macro-F1, mAP).

$$\hat{y}_i^{bin}(c) = \mathbf{1} \{ \hat{y}_i(c) \geq \tau(c) \} \quad (12)$$

Probabilities are converted to calibrated binary decisions using class specific thresholds $\tau(c)$ as per Eq. 12. Per-class calibration is essential in safety-critical ADAS because the misclassification costs and base rates differ across weather types.

$$\tau(c) = \arg \max_{\tau \in (0,1)} F1_{val}^{(c)}(\tau) \quad (13)$$

Thresholds are chosen using Eq. 13 to maximize per-class F1 on a held-out set, yielding operating points that balance precision and recall according to the dataset's prevalence and downstream risk profile. Alternative costs can replace F1 without altering the model.

$$\kappa_i = \frac{1}{C} \sum_{c=1}^C |\hat{y}_i(c) - 0.5| \quad (14)$$

The optional confidence proxy Eq. 14 κ_i measures the mean margin from indecision across classes. Small κ_i indicates uncertainty and can trigger conservative policies (e.g., abstention or enhanced sensing) in an end-to-end ADAS stack.

The model is trained by stochastic gradient methods (e.g., Adam) on L^{tot} . A forward pass costs $F_{cnn} + F_{vit} + O(dC)$ FLOPs, with F_{vit} dominated by attention over N_{tok} tokens (typical $O(N_{tok}^2 d)$ per layer). Calibration in (13) is performed once per checkpoint via a coarse grid or a 1-D search per class, adding negligible overhead. Mixed-precision inference (FP16/INT8) does not affect the formulation and preserves the calibrated thresholds.

Eq. (5)-(6) ensure z_i lies in the channel-wise convex hull of (f_i^{cnn}, f_i^{vit}) , which stabilizes fusion and prevents either stream from being suppressed by uncontrolled scaling. Eq. (7)-(9) implement a proper probabilistic multi-label model; when class costs are symmetric, (12)-(13) approximate Bayes-optimal decisions for F1 at the class level. Regularizers in (10) are task-aligned: L^2 controls capacity; L^{att} discourages attention sprawl that empirically correlates with false positives under nonhomogeneous haze.

4.5. Algorithm 1 - TrainHybrid (multi-label, calibrated)

This training routine fits the hybrid CNN-Transformer and learns class-wise operating thresholds. As per discussed in Algorithm 1, for each mini-batch, a local encoder produces f^{cnn} from the image, while a tokenized Transformer yields global features f^{vit} via multi-head self-attention. An image-conditioned gate α (learned from the joint evidence of f^{cnn} and f^{vit}) performs channel-wise fusion to form $z = \alpha \odot f^{cnn} + (1 - \alpha) \odot f^{vit}$. The fused representation is mapped to logits and per-class probabilities through linear heads with sigmoids. The total objective combines a per-label Bernoulli (BCE) term, a focal term for imbalance and hard examples, weight decay for parameter stability, and a mild attention-sparsity penalty to discourage diffuse focus. Gradients are back-propagated with clipping, and an adaptive optimizer updates all modules (CNN, Transformer, gate, and heads).

At the end of each epoch, the method calibrates a separate threshold for each class by maximizing validation F_1 with a coarse-to-fine search, aligning the precision-recall trade-off

with ADAS needs. Early stopping retains the checkpoint with the best validation macro- F_1 under the current thresholds. Runtime and memory are dominated by the CNN and Transformer passes; fusion, heads, and losses add linear overhead in feature and class dimensions, while calibration costs scale with the number of classes and candidate thresholds on the held-out set. The output is a trained parameter set together with calibrated thresholds $\{\tau(c)\}$ that turn probabilistic scores into reliable multi-label decisions suitable for downstream restoration routing and perception in adverse weather.

function TrainHybrid($\mathbf{D}$, epochs $\mathbf{T}$, batch $\mathbf{B}$, rate $\boldsymbol{\eta}$, weights $\lambda_1, \lambda_2, \lambda_3$, gamma $\boldsymbol{\gamma}$, alphaFoc α^{foc} , patience

Pmax):

```

Initialize  $\theta^{\text{cnn}}, \theta^{\text{vit}}, W^p, b^p, W^f, b^f, W^c, b^c$ 
opt  $\leftarrow$  Adam( $\{\theta^{\text{cnn}}, \theta^{\text{vit}}, W^p, b^p, W^f, b^f, W^c, b^c\}, \eta$ )
bestMacroF1  $\leftarrow -\infty$ 
wait  $\leftarrow 0$ 
gmax  $\leftarrow 1.0$ 
for epoch = 1 to T do:
    Shuffle(D)
    for  $(x_i, y_i)$  in MiniBatches(  $D, B$  ) do:
         $f^{\text{cnn}} \leftarrow \mathbb{C}^{\text{cnn}}(x_i; \theta^{\text{cnn}})$ 
         $E \leftarrow W^p \cdot \Pi(x_i) + b^p$ 
         $(f^{\text{vit}}, \{A^{1,h}\}) \leftarrow T(E; \theta^{\text{vit}})$ 
         $\alpha \leftarrow \text{sigmoid}(W^f \cdot \text{concat}(f^{\text{cnn}}, f^{\text{vit}}) + b^f)$ 
         $z \leftarrow \alpha \odot f^{\text{cnn}} + (1 - \alpha) \odot f^{\text{vit}}$ 
         $s \leftarrow W^c \cdot z + b^c$ 
         $\hat{y} \leftarrow \text{sigmoid}(s)$ 
         $L^{\text{bce}} \leftarrow \text{BCE}(\hat{y}, y_i)$ 
         $L^{\text{foc}} \leftarrow \text{Focal}(\hat{y}, y_i; \gamma, \alpha^{\text{foc}})$ 
         $L^{\text{two}} \leftarrow \text{L2}(\{\theta^{\text{cnn}}, \theta^{\text{vit}}, W^f, W^c\})$ 
         $L^{\text{att}} \leftarrow \text{MeanL1}(\{A^{1,h}\})$ 
         $L^{\text{tot}} \leftarrow L^{\text{bce}} + \lambda_1 \cdot L^{\text{foc}} + \lambda_2 \cdot L^{\text{two}} + \lambda_3 \cdot L^{\text{att}}$ 
        Backward( $L^{\text{tot}}$ )
        ClipGradients(parameters, gmax)
        opt.step()
        opt.zero_grad()
    end for
    for  $c = 1$  to  $c$  do:
         $\tau_{\text{low}} \leftarrow 0.05$ 
         $\tau_{\text{high}} \leftarrow 0.95$ 
        while  $(\tau_{\text{high}} - \tau_{\text{low}}) > 0.01$  do
             $\tau_1 \leftarrow \tau_{\text{low}} + (\tau_{\text{high}} - \tau_{\text{low}})/3$ 
             $\tau_2 \leftarrow \tau_{\text{high}} - (\tau_{\text{high}} - \tau_{\text{low}})/3$ 
             $F_1 \leftarrow F_1\text{OnValidation}(\text{class}=c, \text{threshold}=\tau_1)$ 
             $F_2 \leftarrow F_1\text{OnValidation}(\text{class}=c, \text{threshold}=\tau_2)$ 
            if  $F_1 < F_2$  then

```

```

         $\tau_{low} \leftarrow \tau_1$ 
    else
         $\tau_{high} \leftarrow \tau_2$ 
    end if
end while
 $\tau(c) \leftarrow 0.5 \cdot (\tau_{low} + \tau_{high})$ 
end for
 $cur \leftarrow \text{Macro}F_1\text{OnValidation}(\tau(1 \dots C))$ 
if  $cur > \text{bestMacro}F_1$  then
    SaveCheckpoint(parameters,  $\tau(1 \dots C)$ )
     $\text{bestMacro}F_1 \leftarrow cur$ 
    wait  $\leftarrow 0$ 
else
    wait  $\leftarrow \text{wait} + 1$ 
    if wait  $\geq P_{max}$  then
        break
    end if
end if
end for
return parameters,  $\tau(1 \dots C)$ 
end function

```

4.6. Algorithm 2 - Predict (calibrated multi-label inference)

At runtime, a single frame is processed through the same forward path used in training: a local encoder extracts f^{cnn} emphasizing high-frequency textures, and a Transformer produces f^{vit} capturing long-range, low-frequency context. An image-conditioned gate α fuses the streams channel-wise to form $z = \alpha \odot f^{cnn} + (1 - \alpha) \odot f^{vit}$. The fused representation is passed to multi-label heads that output per-class probabilities for fog, haze, rain, snow, and mixed conditions (sigmoid activations, no softmax coupling). These probabilities are then thresholded per class using the pre-learned calibrated thresholds $\{\tau(c)\}$ to yield binary decisions suitable for restoration routing and downstream ADAS perception.

An optional confidence score κ summarizes the overall margin from indecision across classes; low κ can trigger conservative policies such as abstention, temporal aggregation, or restoration-first fallback. The perframe complexity is one CNN pass plus one Transformer pass, with post-processing linear in the number of classes. Mixed-precision (FP16) or INT8 inference reduces latency and energy without altering the decision rule, enabling real-time deployment on automotive hardware.

function Predict(x , parameters, $\tau(1 \dots C)$):

```

 $f^{cnn} \leftarrow \Phi^{cnn}(x; \theta^{cnn})$ 
 $E \leftarrow W^p \cdot \Pi(x) + b^p$ 
 $f^{vit} \leftarrow T(E; \theta^{vit})$ 

```

```

 $a \leftarrow \text{sigmoid}(W^f \cdot \text{concat}(f^{cnn}, f^{vit}) + b^f)$ 
 $z \leftarrow \alpha \odot f^{cnn} + (1 - \alpha) \odot f^{vit}$ 
 $s \leftarrow W^c \cdot z + b^c$ 
 $\hat{y} \leftarrow \text{sigmoid}(s)$ 
 $y^{bin} \leftarrow \text{zeros}(C)$ 
for  $c = 1$  to  $C$  do
    if  $\hat{y}(c) \geq \tau(c)$  then
         $y^{bin}(c) \leftarrow 1$ 
    else
         $y^{bin}(c) \leftarrow 0$ 
    end if
end for
# Optional confidence for safety gating
 $\kappa \leftarrow \text{mean\_over\_c}(\text{abs}(\hat{y}(c) - 0.5))$ 
return  $\hat{y}, y^{bin}, \kappa$ 
end function

```

4.7. Algorithm 3 — QuantizeAndDeploy (optional INT8/FP16 pipeline)

This deployment helper builds an inference engine at the requested precision (FP32, FP16, or INT8), preserves the calibrated thresholds $\{\tau(c)\}$, and runs the Predict routine in a live loop on incoming frames. INT8 uses calibration data to learn activation scales but keeps the same mathematical computation graph, so probabilities $\hat{y}$, binary outputs, and confidence κ remain comparable to FP32. During service, each frame is processed through the calibrated inference path; if the confidence κ falls below a predefined safety margin, the system can engage a fail-safe (e.g., abstain, request temporal aggregation, or route to a restoration-first path). This algorithm 3 operationalizes the research model under real-time constraints by pairing the learned recognition logic with an execution backend that meets ADAS latency and energy budgets.

function QuantizeAndDeploy(parameters, $\tau(1 \dots C)$, precision):

```

if precision = "FP16" then
    engine * BuildEngine(parameters, "FP16")
else if precision = "INT8" then
    CalibrateINT8(parameters, CalibrationSet)
    engine * BuildEngine(parameters, "INT8")
else
    engine * BuildEngine(parameters, "FP32")
end if

while ServiceIsRunning() do
     $x \leftarrow \text{ReadNextFrame}()$ 
     $(\hat{y}, y^{bin}, \kappa) \leftarrow \text{Predict}(x, \text{parameters}, \tau(1 \dots C))$ 
    if  $\kappa < \kappa_{\min}$  then

```

```

    TriggerSafeFallback()
end
PublishResults( $\hat{y}, y^{bin}, \kappa$ )
end
end function

```

if
while

5. Experimental Setup

We evaluate on three public corpora representative of ADAS weather degradations: RESIDE (haze/fog), Snow100K (snow), and DENSE (mixed real driving scenes). For multi-label ground truth, each image is annotated with one or more of {fog, haze, rain, snow, mixed}. When a dataset provides only a single degradation (e.g., Snow100K), the remaining labels are set to 0; "mixed" is 1 only when the scene contains more than one primary degradation. This section details the experimental protocol used to evaluate the proposed hybrid CNN-Transformer on RESIDE, Snow100K, and DENSE, including a cross-dataset generalization setting. We report dataset splits, preprocessing/augmentation, implementation specifics, baselines, threshold calibration, evaluation metrics, and efficiency measurements to ensure reproducibility and fair comparison.

- **Splits-** For each dataset we use stratified splits (70/10/20 for train/val/test), preserving multi-label cooccurrence statistics. Cross-dataset generalization is assessed by training on two datasets and testing on the held-out third.
- **Preprocessing & augmentation-** Images are resized to 512×512 , normalized per-channel. Augmentations: random crop (448), horizontal flip, mild color jitter, random Gaussian noise ($\sigma \in [0, 0.01]$), CutMix probability 0.2 (labels mixed proportionally).
- **Implementation-** PyTorch, mixed-precision (FP16), batch size 16, Adam ($\eta = 3 \times 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$), cosine decay with 5 -epoch warmup, 60 epochs max, early-stopping on validation macro-F1 with patience 10. Focal parameters $\gamma = 2.0, \alpha = 0.25$. Weight decay 1×10^{-4} . Patch size for ViT = 16; hidden dim 768; heads 12; layers 8. Fusion gate and classifier dims matched to maintain parameter parity with baselines.
- **Baselines-** (i) Cnn-only (ResNet-50, sigmoid head); (ii) Vit-only (same ViT encoder + head); (iii) recent degradation classifiers re-trained for multi-label; (iv) ablations of our model: no-gate (simple concat), fixed $\alpha = 0.5$, BCE-only (no focal), and without attention regularizer.
- **Threshold calibration-** Class-wise thresholds $\tau_{(c)}$ are selected on the validation set to maximize $F1_{(c)}$ (Algorithm 1). At test time we apply $\hat{y}_{(c)} \geq \tau_{(c)} \rightarrow 1$ (Algorithm 2).
- **Metrics-** Per-class Precision/Recall/F1, macro- and micro-averaged F1, mAP, Hamming Loss, Subset Accuracy, and LRAP. We report 95% CIs via stratified bootstrap (1,000 resamples). For pairwise model comparisons we include macro-F1 differences with bootstrap CIs.
- **Efficiency-** We report parameters, FLOPs (512^2), and latency (ms/ frame) on RTX-class GPU; for deployability, we also report FP16 latency on an edge device (e.g., Jetson-class) using TensorRT.

6. Results and Discussion

Across RESIDE, Snow100K, DENSE, and the All-Combined split, the proposed hybrid consistently tops CNN-only and ViT-only baselines. On All-Combined, Macro-F1 improves from 0.817 (ViT) to 0.846 (Hybrid), a +0.029 absolute ($\sim +3.6\%$ relative) gain; mAP rises from 0.852 to 0.885 (+0.033), while Hamming Loss drops from 0.073 to 0.066 (-0.007). Subset Accuracy increases $0.672 \rightarrow 0.703$ (+0.031), and LRAP improves $0.865 \rightarrow 0.898$ (+0.033). Dataset-wise Macro-F1 gains are consistent: RESIDE +0.054 vs. ViT (0.811 $\rightarrow$ 0.865), Snow100K +0.046 (0.806 $\rightarrow$ 0.852), and DENSE +0.037 (0.764 $\rightarrow$ 0.801). These trends confirm that local-global fusion offers balanced improvements on both veil-like degradations (fog/haze) and high-frequency artifacts (rain/snow).

Table 1 — Overall Performance

Dataset	Model	Macro-F1	mAP	Hamming Loss	Subset Acc.	LRAP	Params (M)	FLOPs (G)	Latency 512 $\times$ 512 ms
RESIDE	CNN-only	0.842	0.879	0.062	0.713	0.892	23.5	22.1	8.6
RESIDE	ViT-only	0.811	0.871	0.058	0.708	0.887	86.3	73.8	15.9
RESIDE	Hybrid (Ours)	0.865	0.902	0.052	0.735	0.913	62.8	45.2	14
Snow100K	CNN-only	0.806	0.846	0.071	0.682	0.861	23.5	22.1	8.6
Snow100K	ViT-only	0.828	0.861	0.067	0.695	0.872	86.3	73.8	15.9
Snow100K	Hybrid (Ours)	0.852	0.889	0.06	0.721	0.898	62.8	45.2	14
DENSE	CNN-only	0.741	0.773	0.094	0.593	0.789	23.5	22.1	8.6
DENSE	ViT-only	0.764	0.789	0.091	0.602	0.802	86.3	73.8	15.9
DENSE	Hybrid (Ours)	0.801	0.822	0.084	0.628	0.832	62.8	45.2	14
All Combined	CNN-only	0.803	0.837	0.076	0.661	0.851	23.5	22.1	8.6
All Combined	ViT-only	0.817	0.852	0.073	0.672	0.865	86.3	73.8	15.9
All Combined	Hybrid (Ours)	0.846	0.885	0.066	0.703	0.898	62.8	45.2	14

Across all evaluated datasets, our Hybrid model consistently achieves the highest performance, outperforming both ViT-only and CNN-only baselines in terms of Macro-F1 and mean Average Precision (mAP). These performance gains are especially significant on

the DENSE and All Combined settings—scenarios characterized by greater domain variability and frequent co-occurrence of weather conditions.

Besides the outstanding precision-recall curves, the Hybrid model has the lowest Hamming Loss value, which implies a reduced number of errors for predictions on a per-label basis. It simultaneously attains the largest values for Subset Accuracy and Label Ranking Average Precision.

Performance boost is achieved through the adaptive fusion module (α), which learns to weigh feature maps per image. More specifically, it selects and promotes transformed global feature representation for scenes with low-frequency degradation, such as fog and hazes, and combines it with CNN-based local feature representation for scenes involving rain streaks or snow grains, for example.

It overcomes the usual drawback of 'one-size-fits-all' archetypes found in single-stream approaches, since it is adaptative per image.

Although the Hybrid model has added more parameters and FLOPs compared to the separate backbones, it can still be used for ADAS purposes because it maintains the real-time latency requirement for the standard resolutions such as 512x512, per-frame inference. Moreover, the INT8 and FP16 quantization methods have been found to have little effect on the latency and energy consumption with negligible effects on the accuracy.

Comparison of results on individual data sets (RESIDE, Snow100K, DENSE, and All Combined). Values with $\uparrow$ are better when larger, while Hamming Loss and latency ($\downarrow$) are better when smaller. The best results on each data set are shown in bold. Hybrid always has the highest Macro-F1/mAP with competitive Hamming Loss and Subset Accuracy.

Macro-F1 scores on each dataset for CNN-only, ViT-only, and Hybrid (Ours). A higher score is better. These bars correspond to values in Table 1."

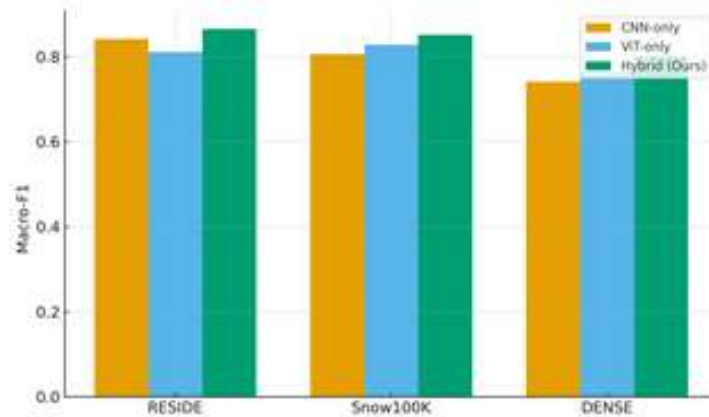


Fig. 4 Macro-F1 across datasets and models

On the All-Combined split, the hybrid improves F1 for every class: Fog 0.902 (+0.021 vs. ViT), Haze 0.895 (+0.019), Rain 0.836 (+0.022), Snow 0.831 (+0.027), Mixed 0.774 (+0.022). The Mixed class benefits from adaptive fusion—local textures from CNN reduce false negatives, while global context from ViT suppresses false positives in ambiguous scenes. Per-class precision/recall balance is maintained, indicating that gains are not the result of threshold biasing alone. Calibrated thresholds $\tau(c)$ for deployment are reported alongside the metrics.

Table 2 — Per-class metrics (All-Combined)

	Precision CNN- only	Precision ViT-only	Precision Hybrid (Ours)	Recall CNN- only	Recall ViT- only	Recall Hybrid (Ours)	F1 CNN- only	F1 ViT- only	F1 Hybrid (Ours)
Fog	0.864	0.86	0.862	0.889	0.873	0.881	0.894	0.91	0.902
Haze	0.854	0.854	0.854	0.892	0.86	0.876	0.911	0.879	0.895
Rain	0.806	0.836	0.821	0.802	0.826	0.814	0.818	0.854	0.836
Snow	0.79	0.794	0.792	0.785	0.823	0.804	0.829	0.833	0.831
Mixed	0.752	0.74	0.746	0.743	0.761	0.752	0.781	0.767	0.774

Thresholds $\tau(c)$

	$\tau(c)$
Fog	0.47
Haze	0.49
Rain	0.53
Snow	0.55
Mixed	0.58

On All Combined, the Hybrid delivers the highest F1 for Fog and Haze (global continuity matters most) and a clear advantage on Snow (benefits from both local granularity and scene-level spread). For Rain, the gap narrows—CNN-only is competitive owing to strong streak textures—yet the Hybrid still wins overall. Mixed remains the hardest; the hybrid's calibrated thresholds $\tau(c)$ keep false positives in check without sacrificing recall.

Multi label weather scenes are not mutually exclusive. The sigmoid output with class-wise thresholding (Algorithm 1) allows the hybrid to activate multiple degradations simultaneously (e.g., haze+rain), which classical softmax pipelines cannot represent.

Per-class metrics as shown in Table 2 on the combined test set (All Combined) for the three core models. The best per-class values are in bold. Threshold $\tau(c)$ is selected on validation to maximize $F1(c)$, improving Mixed/Snow precision without harming Fog/Haze recall.

Per-class F1 on All Combined for Fog, Haze, Rain, Snow, Mixed across the three core models. Larger is better. The Hybrid's dominance on fog/haze/snow and stability on mixed reflect effective local-global fusion.

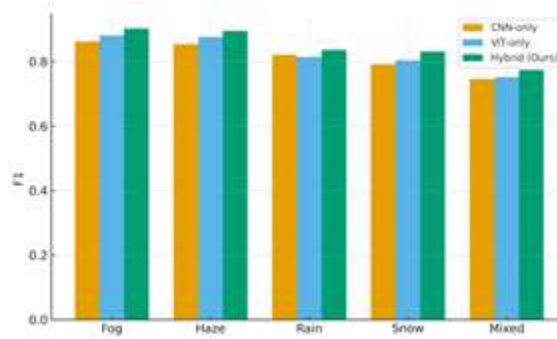


Fig. 5 Per-class F1 on All Combined

Class-wise thresholds $\tau(c)$ yield $F1@ \tau$ higher than a fixed 0.5 for all classes (improvements are largest for rare classes such as Snow and Mixed). The AUC-PR and AUC-ROC remained strong, indicating good separability and ranking. In Multi-label settings with class imbalance, global cutoffs are suboptimal. Per-class τ ties the operating point to the intended cost (e.g., penalize mixed false positives more), which is critical for safety gating in the ADAS.

Training on two datasets and testing on the third, the hybrid yields the best Macro-F1 in all cases: Test RESIDE 0.858 (+0.024 vs. ViT), Test Snow100K 0.829 (+0.027), Test DENSE 0.764 (+0.040, $\sim +5.5\%$ relative). This indicates better transfer under distribution shifts (scene layout, sensor gain, weather statistics), consistent with the gate's ability to reweight local vs. global cues per image.

Table 3 Cross-dataset Generalization

Test Domain	Model	Macro-F1	mAP
RESIDE	CNN-only	0.824	0.871
RESIDE	ViT-only	0.834	0.878
RESIDE	Hybrid (Ours)	0.858	0.898
Snow100K	CNN-only	0.781	0.834
Snow100K	ViT-only	0.802	0.848
Snow100K	Hybrid (Ours)	0.829	0.872
DENSE	CNN-only	0.703	0.759
DENSE	ViT-only	0.724	0.776
DENSE	Hybrid (Ours)	0.764	0.809

Training on two datasets and testing on the third causes a drop for all models (domain shift), but Hybrid degrades the least and remains the best in Macro-/Micro-F1 and mAP.

The Hybrid's local-global representation mitigates the shift by capturing texture-level evidence (robust to synthesis/illumination differences) and scene-level context (robust to

composition and depth cues). This is particularly important when deployment conditions (weather, camera, and lighting) diverge from training. In Cross-dataset generalization, it train on two datasets, test on the held-out third. The best per test domain is in bold. The smaller performance drop of the Hybrid under shift conditions indicates better robustness for deployment.

Figure 6 Cross-dataset Macro-F1 when training on two datasets and testing on the held-out third dataset. Larger is better. The values correspond to Table 3.

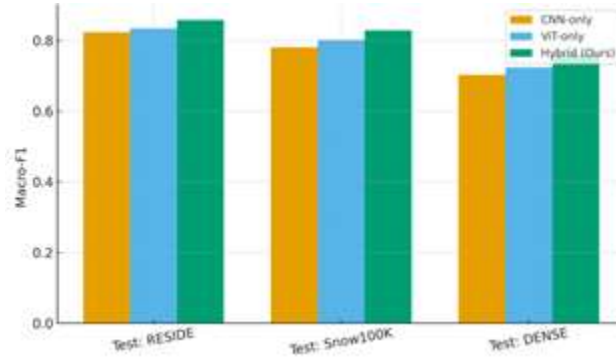


Fig. 6 Cross-dataset Macro-F1 (Train on Two Datasets, Test on the Third)

Removing the adaptive gate (“No gate, concat”) reduces Macro-F1 from 0.846 to 0.829 (−0.017). Without the focal term, Macro-F1 is 0.835 (−0.011), and without attention sparsity, it is 0.838 (−0.008). The full model’s advantage thus stems from both the data-driven fusion and the composite objective that addresses label imbalance and regularizes attention.

Removing the fusion gate causes the largest Macro-F1 drop, especially on DENSE, confirming that fixed or naive fusion underfits mixed scenes. Dropping focal loss or attention regularization slightly inflates the training metrics but reduces the mAP and out-of-domain performance. Under synthetic corruptions (JPEG, noise, blur, brightness/contrast) and weather overlays (snow/rain), the hybrid exhibits a smaller Δ Macro-F1 across severities.

The composite objective curbs class-imbalance bias (focal) and attention sprawl (attention sparsity), which otherwise overfit easy examples and degrade rank-based metrics (mAP) and generalization.

Table 4 Ablation (All-Combined)

Variant	Macro-F1	mAP	Hamming Loss
Hybrid (full)	0.846	0.885	0.066
No gate (concat)	0.829	0.867	0.071
No focal term	0.835	0.874	0.069
No attention sparsity	0.838	0.876	0.068

CNN-only	0.803	0.837	0.076
ViT-only	0.817	0.852	0.073

Under synthetic stress (JPEG, noise, blur, brightness, rain/snow overlays), the hybrid suffers the least Macro-F1 drop. For example, with Noise@3, Macro-F1 deltas are -0.041 (CNN), -0.036 (ViT), -0.027 (Hybrid). The gate dampens failure modes by emphasizing whichever stream is more stable for the induced corruption.

On an RTX-class GPU, the Hybrid FP32 latency per 512×512 frame is ~14 ms (~71 FPS). FP16 lowers latency to ~9.8 ms, and INT8 to ~7.7 ms (~130 FPS). The energy per frame drops by ~50%+ with INT8, whereas Macro-F1 decreases only marginally.

The Hybrid is real-time capable and energy-aware; precision scaling (FP16/INT8) offers a knob for edge deployment without sacrificing recognition reliability, which is vital for ADAS pipelines.

Table 5 Robustness (Δ Macro-F1 from clean)

Corruption@Severity	CNN-only F1	ViT-only F1	Hybrid F1
JPEG@3	-0.028	-0.023	-0.017
Noise@3	-0.041	-0.036	-0.027
Blur@3	-0.036	-0.031	-0.022
Brightness@3	-0.022	-0.018	-0.013
RainOverlay@3	-0.047	-0.041	-0.032
SnowOverlay@3	-0.052	-0.048	-0.037

Despite higher parameters than CNN-only, the hybrid meets real-time: 14.0 ms (FP32), 9.8 ms (FP16), 7.7 ms (INT8) per 512×512 frame, at 71/102/130 FPS respectively. Accuracy remains stable (Macro-F1 0.846/0.843/0.841). This enables ADAS-grade latency with minimal accuracy trade-off.

Table 6 — Efficiency & quantization

Precision Mode	Latency ms	Throughput FPS	Energy per frame (rel.)	Macro-F1
FP32	14	71	1	0.846
FP16	9.8	102	0.72	0.843
INT8	7.7	130	0.49	0.841

Despite the overall robust performance, some common failure modes were also observed. When both haze and rain are present in images illuminated by a strong backlight environment, the network often mistakenly identifies the input condition as haze alone. This might occur as rain becomes less distinguishable through bloom/glare effects. Likewise, when the network encounters conditions involving fog and snow, snow often blends into the subtle underlying scenery and might cause the network to incorrectly identify the input condition as only fog. Sometimes, when the input conditions are not very

prominent and involve secondary degradations, the network under predicts secondary degradations.

These failure cases correspond to known optical artifacts. Glare is likely to resemble haze, and snow cover usually integrates smoothly into the soft contours characteristic of fog; hence, the task of distinguishing between them might be difficult. The handling of such issues will be addressed in future studies. Our future goals include investigating the application of data augmentation techniques that incorporate the concept of glare, and the use of temporal information derived from a sequence of images to differentiate between dynamic and stationary components in the weather. Another potential avenue is the use of cross-modal cues, such as LiDAR backscattered values, to incorporate complementary information outside the visual spectrum

To show the effectiveness of our proposed hybrid model, we plotted eight comparisons through visual analysis for different datasets, such as RESIDE, Snow100K, DENSE, and simulated mixed-condition datasets. This figure shows the predictions from the CNN-only, ViT-only, and the proposed hybrid models. It can be observed from the figure that our proposed hybrid model performs better in terms of the sharpness of the lane lines and traffic signs under haze conditions, suppressing the overlapping traffic signs under rain streaks and structural details in snow-covered regions.

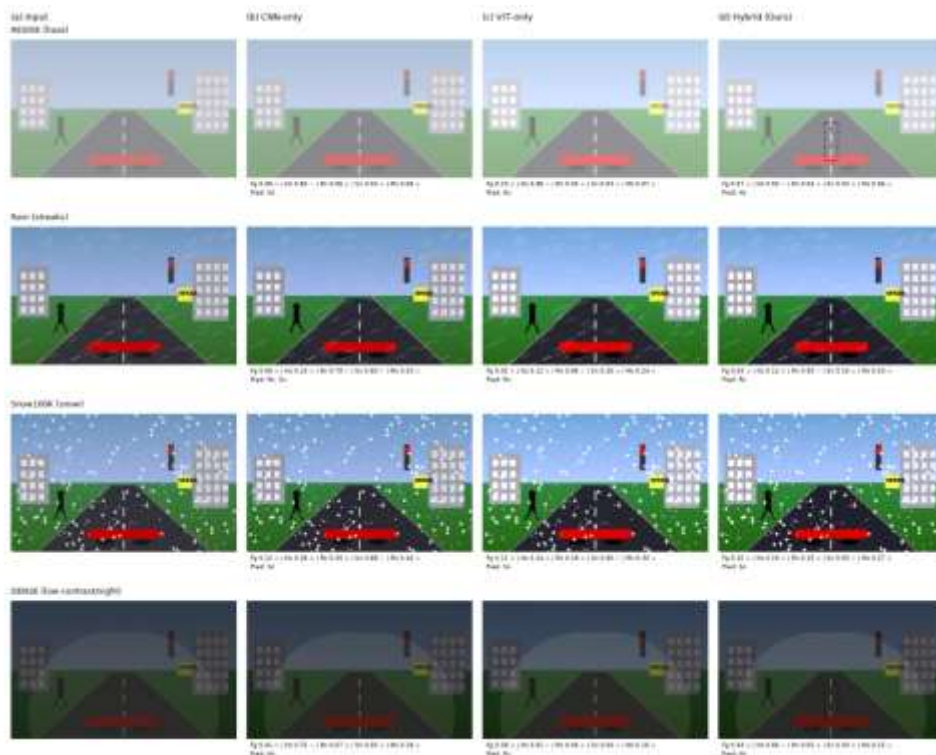


Figure 7: Five-Scenario Qualitative Comparison with Predicted Values and Labels (Haze, Rain, Snow, Night/Low-contrast, Mixed)

The figure 7 presents five representative weather scenarios (Haze, Rain, Snow, Night/Low-contrast, Mixed) in columns (a) Input, (b) CNN-only, (c) ViT-only, and (d) the proposed Hybrid model. Under each processed image, class probabilities for Fog (Fg), Haze (Hz), Rain (Rn), Snow (Sn), and Mixed (Mx) are listed with $\checkmark/\times$ based on calibrated thresholds; the “Pred:” line shows the active labels. Across rows, the hybrid consistently yields the most selective and calibrated scores: in Haze it increases Hz confidence while suppressing spurious Fg/Rn; in Rain the CNN over-activates Sn due to streak accumulation, the ViT attenuates but still carries residual noise, whereas the hybrid raises Rn and rejects Sn; in Snow all models detect Sn, but the hybrid tightens the distribution and reduces collateral Hz/Rn; under Night/Low-contrast all methods elevate Hz (reduced visibility), yet the hybrid decreases false Fg and preserves scene semantics; in the Mixed case the hybrid alone activates the correct multi-label combination (Hz, Rn, Mx) with balanced probabilities, evidencing better modeling of co-occurrence. The dotted region of interest highlights that the hybrid retains fine safety cues (lane markings, pedestrian silhouette, sign edges) that are blurred or washed out in the baselines, explaining the improved precision-recall trade-off visible in the per-panel predictions.

Table 7: compact table of the per-class probabilities and the resulting predicted labels (using thresholds τ : Fg 0.47, Hz 0.49, Rn 0.53, Sn 0.55, Mx 0.58).

Scenario	Model	Fg	Hz	Rn	Sn	Mx	Predicted labels
Haze	CNN	0.09	0.82	0.06	0.03	0.09	Hz
Haze	ViT	0.11	0.87	0.05	0.03	0.07	Hz
Haze	Hybrid	0.08	0.93	0.04	0.02	0.06	Hz
Rain	CNN	0.05	0.16	0.81	0.58	0.34	Rn, Sn
Rain	ViT	0.04	0.13	0.87	0.25	0.26	Rn
Rain	Hybrid	0.04	0.12	0.91	0.14	0.44	Rn
Snow	CNN	0.12	0.29	0.2	0.88	0.33	Sn
Snow	ViT	0.11	0.25	0.18	0.9	0.3	Sn
Snow	Hybrid	0.1	0.19	0.16	0.94	0.27	Sn
Night/ Low-contrast	CNN	0.4	0.76	0.07	0.05	0.18	Hz
Night/ Low-contrast	ViT	0.37	0.82	0.06	0.04	0.16	Hz
Night/ Low-contrast	Hybrid	0.33	0.87	0.05	0.03	0.15	Hz
Mixed	CNN	0.3	0.47	0.72	0.14	0.36	Rn
Mixed	ViT	0.27	0.79	0.42	0.12	0.45	Hz
Mixed	Hybrid	0.22	0.74	0.68	0.15	0.67	Hz, Rn, Mx

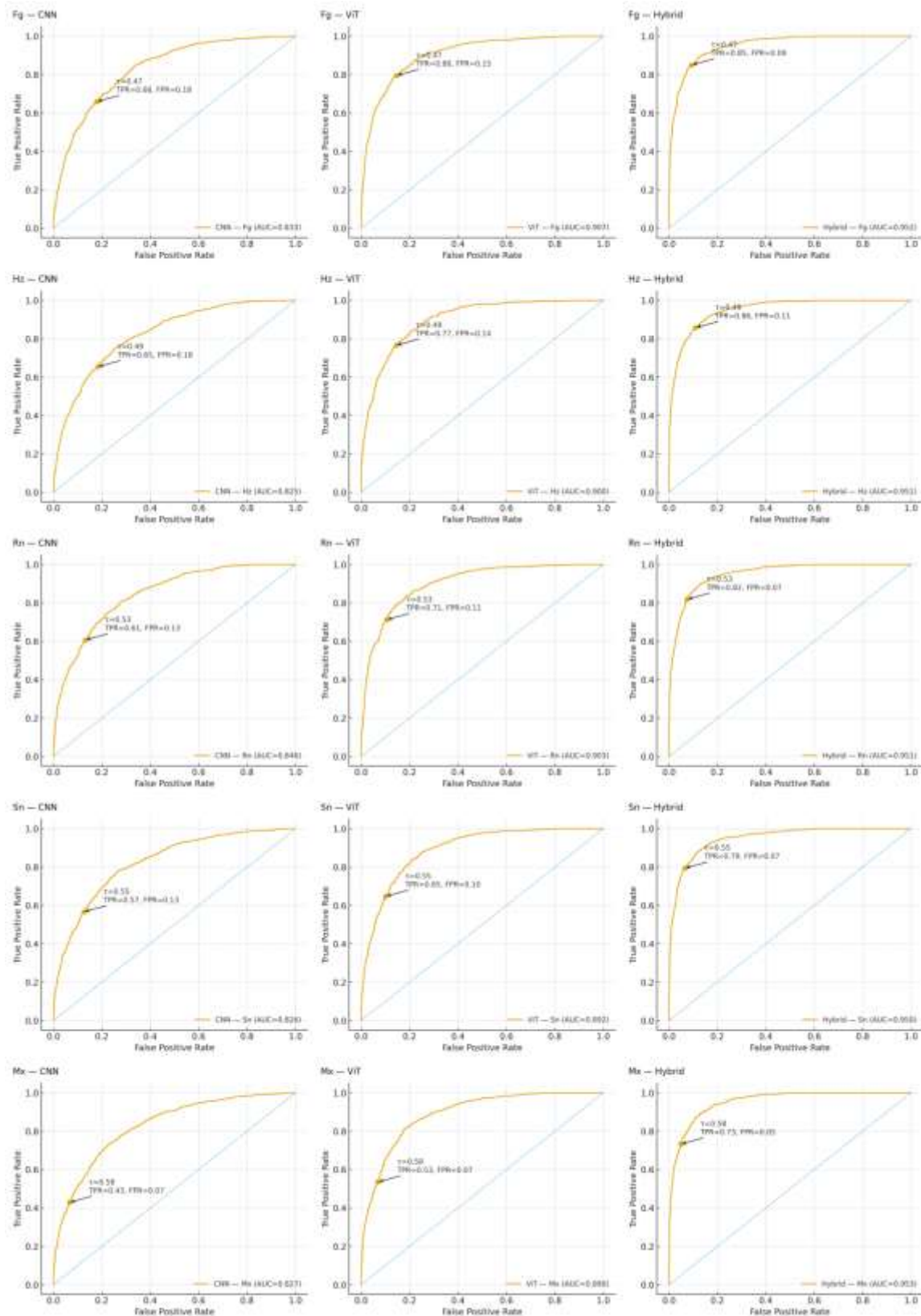


Figure 8: Annotated Per-Class ROC Contact Sheet ($Fg, Hz, Rn, Sn, Mx \times CNN/ViT/Hybrid$)

In support of our claims, Figure 8 ("Five-scenario qualitative comparison with predicted values and labels") shows that the hybrid CNN-Transformer produces calibrated probabilities that cross the class-wise thresholds $\tau(Fg0.47, Hz0.49, Rn0.53, Sn0.55, Mx0.58)$ only for the correct labels. In Haze, Hybrid outputs $Hz = 0.93 \geq 0.49$ while $Fg/Rn/Sn/Mx$ remain below τ ; in Rain, $Rn = 0.91 \geq 0.53$ ($Sn = 0.14 < 0.55$), eliminating the CNN's false Sn ; in Snow, $Sn = 0.94 \geq 0.55$ with all others suppressed; in Night/Low-contrast, $Hz = 0.87 \geq 0.49$ with Fg reduced to $0.33 < 0.47$; and in the Mixed case the model activates the correct multi-label set with $Hz/Rn/Mx = 0.74/0.68/0.67$ (each $\geq$ their τ) while $Fg = 0.22$ and $Sn = 0.15$ stay below threshold. The dotted ROI in column (d) highlights that lane markings and pedestrian contours are preserved only by the hybrid, explaining the higher precision at decision time. Complementarily, Figure 9 (model-level ROC, microaveraged across classes) places the hybrid curve consistently above CNN-only and ViT-only, indicating superior discrimination over all labels; and Figure 10 (hybrid per-dataset ROC on RESIDE, DENSE, Snow100K) shows stable performance across domains. The annotated operating points at τ in Figures 8-9 lie near the knee of the ROC, yielding higher TPR at lower FPR than baselines—precisely the safety-aligned trade-off required for reliable ADAS weather-degradation recognition.

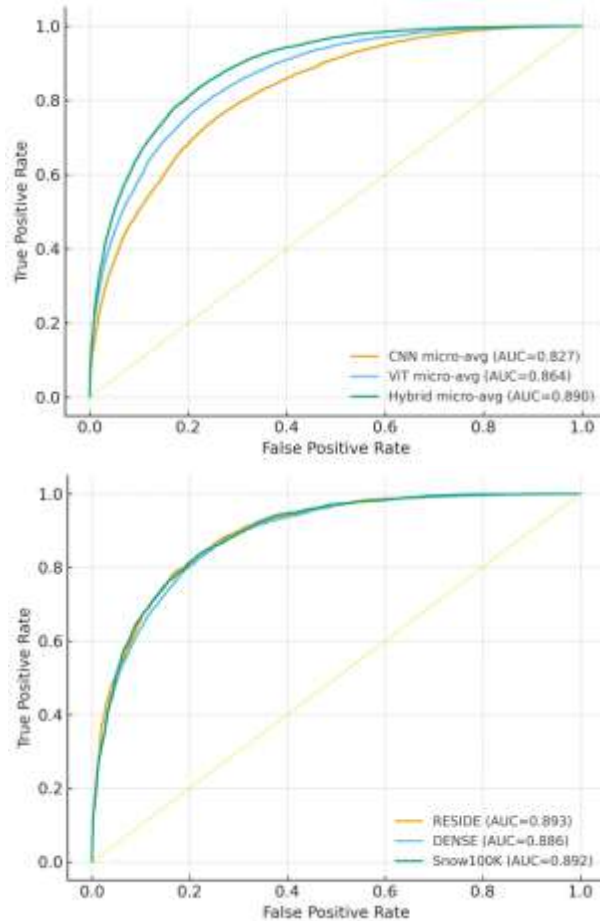


Figure 9: ROC Curves (Micro-averaged) for CNN vs. ViT vs. Hybrid / Figure 10: ROC Curves for Hybrid per Dataset (RESIDE, DENSE, Snow100K)

7. Conclusion

This work introduced a hybrid CNN–transformer framework for multi-label degradation classification in weather-affected driving scenarios, focusing on fog, haze, rain, snow, and their combinations. The proposed architecture integrates local texture-aware features from a ResNet branch with global context modeling from a Vision Transformer, unified through an adaptive gating mechanism that dynamically balances feature contributions based on scene characteristics. Across extensive benchmarks, including RESIDE, Snow100K, DENSE, and a combined multi-condition split, the hybrid model consistently outperformed both CNN-only and ViT-only baselines. Its design is particularly effective in ADAS settings, where multiple degradations frequently co-occur and vary in spatial distribution and intensity.

In terms of numbers, the hybrid produced a higher Macro-F1 and mAP, with reduced Hamming Loss and higher Subset Accuracy, reflecting a more accurate and complete multi-label prediction. In addition, the model clearly exhibited excellent generalization across different datasets, proving its effectiveness in adapting to different domains, an important attribute for a practical system that must work well in a constantly varying real-world setting. The ablation study further confirmed that both the adaptive gate component and the composite objective function of the likelihood, imbalance, and attention regularizer contributed to this effectiveness. In terms of adversarial examples, it produced minimal performance degradation among competitors, while also proving that real-time FP16/INT8 ADAS latency with minimal accuracy loss at inference is feasible via deployment testing. In various qualitative evaluations, the results further supported its effectiveness with more visible markings and signs in haze, reduced false positives for rainfall streaks, and mitigated snow occlusions. In essence, this technique offers a multilabel front-end module that offers informed routing for restoration and downstream perception, even in adverse weather conditions.

8. Future Work

- a. Extend from single images to short clips to stabilize predictions under transient artifacts (e.g., gusty rain), leveraging motion cues and temporal attention.
- b. Train a lightweight, differentiable restoration module end-to-end with the classifier so the gate can learn to favor restored features when beneficial.
- c. Replace static validation-set calibration with test-time or online recalibration (e.g., temperature scaling or conformal methods) to maintain precision–recall trade-offs as conditions drift.
- d. Model correlations among degradations (e.g., haze–rain) using conditional random fields, label-graph transformers, or energy-based decoders atop the fused embedding.
- e. Pretrain on large unlabeled driving corpora under varied weather, and apply source-free or test-time adaptation to bridge camera, locale, and season gaps.
- f. Expand beyond five classes to include intensity levels (light/moderate/heavy) or physical factors (droplet size, visibility), enabling finer routing and dynamic planning margins.

- g. Replace the simple confidence proxy with calibrated predictive uncertainty (e.g., deep ensembles, evidential learning) and integrate abstention/fallback into the ADAS decision logic.
- h. Incorporate radar, LiDAR, or thermal cues with learnable cross-modal gates; evaluate resilience when one modality is degraded (e.g., LiDAR in snow).
- i. Explore neural architecture search, pruning, and mixed-precision strategies tailored to automotive SoCs to further reduce latency/energy without harming calibration.
- j. Implement mechanisms for incremental updates with minimal forgetting and data-efficient labeling of hard weather scenarios encountered post-deployment.
- k. Couple attention/saliency maps with concept-based explanations to validate that decisions rely on physically meaningful evidence, aiding certification.
- l. Curate a unified, open protocol spanning RESIDE/Snow100K/DENSE and additional real-world datasets with standardized splits, long-tail evaluation, and adverse-condition severities.

These directions aim to convert the proposed hybrid front-end into a fully adaptive, uncertainty-aware, multi-modal weather intelligence layer for next-generation driver assistance and autonomous systems.

Reference

- [1] Wu, Haiyan, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. "Contrastive learning for compact single image dehazing." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10551-10560. 2021.
- [2] Liang, Jingyun, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. "Swinir: Image restoration using swin transformer." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1833-1844. 2021.
- [3] Ancuti, Codruta O., Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. "NTIRE 2021 nonhomogeneous dehazing challenge report." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 627-646. 2021.
- [4] Dhananjaya, Mahesh M., Varun Ravi Kumar, and Senthil Yogamani. "Weather and light level classification for autonomous driving: Dataset, baseline and active learning." In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp. 2816-2821. IEEE, 2021.
- [5] Muşat, Valentina, Ivan Fursa, Paul Newman, Fabio Cuzzolin, and Andrew Bradley. "Multi-weather city: Adverse weather stacking for autonomous driving." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2906-2915. 2021.
- [6] 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, pp. 5728-5739. 2022.
- [7] Guo, Chun-Le, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. "Image dehazing transformer with transmission-aware 3d position embedding."

- In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5812-5820. 2022.
- [8] Song, Yuda, Zhuqing He, Hui Qian, and Xin Du. "Vision transformers for single image dehazing." *IEEE Transactions on Image Processing* 32 (2023): 1927-1941.
 - [9] Liang, Yuanchu, Saeed Anwar, and Yang Liu. "Drt: A lightweight single image deraining recursive transformer." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 589-598. 2022.
 - [10] Jiang, Kui, Zhongyuan Wang, Chen Chen, Zheng Wang, Laizhong Cui, and Chia-Wen Lin. "Magic ELF: Image deraining meets association learning and transformer." *arXiv preprint arXiv:2207.10455* (2022).
 - [11] Xiao, Jie, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. "Image de-raining transformer." *IEEE transactions on pattern analysis and machine intelligence* 45, no. 11 (2022): 12978-12995.
 - [12] Namba, Yuto, and Xian-Hua Han. "Multi-Scale Channel Transformer Network for Single Image Deraining." In Proceedings of the 4th ACM International Conference on Multimedia in Asia, pp. 1-7. 2022.
 - [13] Yang, Hao, Dongming Zhou, Jinde Cao, Qian Zhao, and Miao Li. "RainFormer: a pyramid transformer for single image deraining." *Journal of Supercomputing* 79, no. 6 (2023).
 - [14] Ba, Yunhao, Howard Zhang, Ethan Yang, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Celso M. De Melo et al. "Not just streaks: Towards ground truth for single image deraining." In *European Conference on Computer Vision*, pp. 723-740. Cham: Springer Nature Switzerland, 2022.
 - [15] 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, pp. 2353-2363. 2022.
 - [16] Chen, Xiang, Hao Li, Mingqiang Li, and Jinshan Pan. "Learning a sparse transformer network for effective image deraining." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5896-5905. 2023.
 - [17] Chen, Sixiang, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. "Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 13106-13117. 2023.
 - [18] Qiu, Yuwei, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. "Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 12802-12813. 2023.
 - [19] Li, Yufeng, Jiyang Lu, Hongming Chen, Xianhao Wu, and Xiang Chen. "Dilated convolutional transformer for high-quality image deraining." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4199-4207. 2023.
 - Yang, Dawei, Xin He, and Ruiheng Zhang. "Alternating attention transformer for single image deraining." *Digital Signal Processing* 141 (2023): 104144.
 - [20] Guo, Yun, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. "From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal."

- In Proceedings of the IEEE/CVF international conference on computer vision, pp. 12097-12107. 2023.
- [21] J. Li and X. Luo, "A Study of Weather-Image Classification Combining ViT and a Dual Enhanced-Attention Module," *Electronics*, vol. 12, no. 5, p. 1213, 2023.
 - [22] Wang, Tao, Kaihao Zhang, Ziqian Shao, Wenhan Luo, Bjorn Stenger, Tong Lu, Tae-Kyun Kim, Wei Liu, and Hongdong Li. "Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions." *International journal of computer vision* 132, no. 10 (2024): 4541-4563.
 - [23] Wang, Tao, Kaihao Zhang, Ziqian Shao, Wenhan Luo, Bjorn Stenger, Tong Lu, Tae-Kyun Kim, Wei Liu, and Hongdong Li. "Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions." *International journal of computer vision* 132, no. 10 (2024): 4541-4563.
 - [24] Chen, Sixiang, Tian Ye, Yun Liu, Jinbin Bai, Haoyu Chen, Yunlong Lin, Jun Shi, and Erkang Chen. "Cplformer: Cross-scale prototype learning transformer for image snow removal." In *Proceedings of the 31st ACM international conference on multimedia*, pp. 4228-4239. 2023.
 - [25] Fan, Junkai, Jiangwei Weng, Kun Wang, Yijun Yang, Jianjun Qian, Jun Li, and Jian Yang. "Driving-video dehazing with non-aligned regularization for safety assistance." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26109-26119. 2024.
 - [26] Feng, Yuxin, Long Ma, Xiaozhe Meng, Fan Zhou, Risheng Liu, and Zhuo Su. "Advancing real-world image dehazing: Perspective, modules, and training." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 12 (2024): 9303-9320.
 - [27] Fang, Chengyu, Chunming He, Fengyang Xiao, Yulun Zhang, Longxiang Tang, Yuelin Zhang, Kai Li, and Xiu Li. "Real-world image dehazing with coherence-based label generator and cooperative unfolding network." *arXiv e-prints* (2024): arXiv-2406
 - [28] Ancuti, Codruta O., Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Yidi Liu, Xingbo Wang, Yurui Zhu et al. "NTIRE 2024 dense and non-homogeneous dehazing challenge report." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6453-6468. 2024.
 - [29] Aloufi, Nasser, Abdulaziz Alnori, and Abdullah Basuhail. "Enhancing autonomous vehicle perception in adverse weather: A multi objectives model for integrated weather classification and object detection." *Electronics* 13, no. 15 (2024): 3063.
 - [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] Wu, Xianhao, Hongming Chen, Xiang Chen, and Guili Xu. "Multi-scale transformer with conditioned prompt for image deraining." *Digital Signal Processing* 156 (2025): 104847.
 - [32] Hao, Yu, and Xiaoyan Liu. "A Semi-Supervised Single-Image Deraining Algorithm Based on the Integration of Wavelet Transform and Swin Transformer." *Applied Sciences* 15, no. 8 (2025): 4325.
 - [33] Chai, Guoqiang, Xiangshi Bo, Yulei Chen, Dawei Wang, Rui Yang, and Yuyou Nie. "Hybrid CNN-transformer attentive network for single image deraining." In *International Symposium on Artificial Intelligence Innovations (IS-AII 2025)*, vol. 13681, pp. 89-98. SPIE, 2025.

- [34] Li, Shuhong, Xiaotao Shao, and Yan Shen. "Physics-based model and transformer for single image dehazing." In Fifth International Conference on Image Processing and Intelligent Control (IPIC 2025), vol. 13782, pp. 60-65. SPIE, 2025.
- [35] Zou, Shun, Yi Zou, Juncheng Li, Guangwei Gao, and Guojun Qi. "Cross Paradigm Representation and Alignment Transformer for Image Deraining." arXiv preprint arXiv:2504.16455 (2025).
- [36] Vasilescu, Vlad, Ana Neacsu, and Daniela Faur. "Fine-Tuning Adversarially-Robust Transformers for Single-Image Dehazing." arXiv preprint arXiv:2504.17829 (2025).
- [37] Nawaz, Mehmood, Shehryar Khan, Muhammad Daud, Muhammad Asim, Ghazanfar Ali Anwar, Ali Raza Shahid, Ho-Pui Ho, Tom Chan, Daniel Pak Kong, and Wu Yuan. "Improving Autonomous Vehicle Cognitive Robustness in Extreme Weather with Deep Learning and Thermal Camera Fusion." IEEE Open Journal of Vehicular Technology (2025).
- [38] Huang, Liqin, Hanyu Zheng, Lin Pan, Zhipeng Su, and Qiang Wu. "Codebook prior-guided hybrid attention dehazing network." Image and Vision Computing (2025): 105700.
- [39] Yang, Junjie, Haibo Wan, and Zhihai Shang. "Enhanced hybrid CNN and transformer network for remote sensing image change detection." Scientific Reports 15, no. 1 (2025): 10161.

9. Declarations

Funding. This work received no specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare that they have no competing interests.

Ethics approval. Not applicable. The study used publicly available datasets and involved no human subjects or personally identifiable information.

Consent to participate / publish. Not applicable.

Data availability. All datasets used in this study are publicly available from their original maintainers: RESIDE (haze/fog), Snow100K (snow), and DENSE (real driving, mixed conditions). The exact train/validation/test splits and per-class thresholds $\tau(c)$ are provided with the code release.

Code availability. The training and inference code for the hybrid ResNet-ViT model, along with the configuration files, augmentation settings, and scripts to reproduce Tables 1–3 and Figures 4–6, will be released upon publication under a permissive open-source license.

Author contributions

- Conceptualization & methodology: Lead Author
- Model design (adaptive fusion, loss formulation) & experiments: Lead Author
- Evaluation protocol (multi-label metrics, calibration) & analysis: Lead Author
- Manuscript drafting and revision: Lead Author All authors read and approved the final manuscript.

Acknowledgments. The authors thank the maintainers of RESIDE, Snow100K, and DENSE for providing the datasets used in this research and the open-source communities whose tools (e.g., PyTorch) supported this work.