Restoring Vision Under Transitional Weather: A Benchmark for Autonomous Driving Perception

Anonymous authors
Paper under double-blind review

Abstract

Adverse weather conditions substantially degrade autonomous vehicle (AV) perception, with rain, fog, and snow causing severe drops in detection and segmentation accuracy. The challenge is amplified under transitional weather, where conditions evolve dynamically (e.g., foggy to rainy, snowy to foggy), introducing non-stationary domain shifts and temporally varying degradations that hinder both restoration and high-level perception. Existing restoration methods are limited to single, discrete weather types, while recent multi-weather approaches rely on multi-encoder designs that incur high computational costs and struggle to generalize to transitional conditions, thereby limiting scalability and real-time deployment. To address these limitations, we propose TransWeatherNet, a modular two-stage framework for restoring transitional weather to enhance AV perception. At its core, we introduce TWRFormer, a transformer-based restoration network that jointly models global spatiotemporal dependencies and localized degradations through an adaptive encoder-decoder with learned weather embeddings and modified perceptual loss, while employing a lightweight depthwise separable projection head for image reconstruction. Using restored images with gains of +2.01 *dB* PSNR and +0.03 SSIM, TransWeatherNet boosts downstream perception, achieving improvements of +3.1% mAP, +3.5% IoU, and +0.65% BLEU across object detection, semantic and instance segmentation, image captioning, and visual question answering. Extensive experiments validate state-of-the-art restoration and perception performance on AIWD16, Foggy Cityscapes, RainCityscapes, ACDC, and BDD100K.

1 Introduction

Autonomous vehicle (AV) technology is advancing rapidly, with vehicles increasingly being deployed in real-world driving scenarios as we move closer to full autonomy (Liao et al., 2025; Li et al., 2025a; Oksuz et al., 2026; Yu et al., 2026). However, adverse weather remains a persistent challenge that significantly degrades AV perception accuracy, control stability, or navigation robustness (Galshetwar et al., 2026; Zheng et al., 2026). Discrete weather conditions such as rain, fog, snow, and clouds impair camera sensors by reducing visibility and introducing noise in the captured scenes (Yang et al., 2022; Leroux et al., 2022). This degradation in visual fidelity compromises scene understanding, increases the risk of incorrect perception, and ultimately poses a threat to the safety of road users (Lin et al., 2026; Li et al., 2026). Moreover, perception tasks, such as scene classification, object detection, segmentation, image captioning, and visual question answering (VQA), perform reliably under clear daylight conditions but experience substantial degradation in adverse weather, primarily due to reduced visibility and distorted scene content (Sun et al., 2025; Yang et al., 2026; Kondapally et al., 2024b).

Therefore, the development of robust image restoration techniques capable of recovering clean, high-fidelity images from weather-corrupted observations has emerged as a critical research imperative (Liu et al., 2026; Rajagopalan & Patel, 2025). The significance of this endeavor is underscored by several compelling factors. *(i)* In safety-critical domains such as autonomous navigation, and intelligent transportation systems, restored visual clarity directly translates to more accurate perception, enabling timely and informed decision-making under transitional weather conditions. *(ii)* For real-time surveillance and monitoring applications, effective weather removal substantially enhances situational awareness by improving the detectability of obstacles,

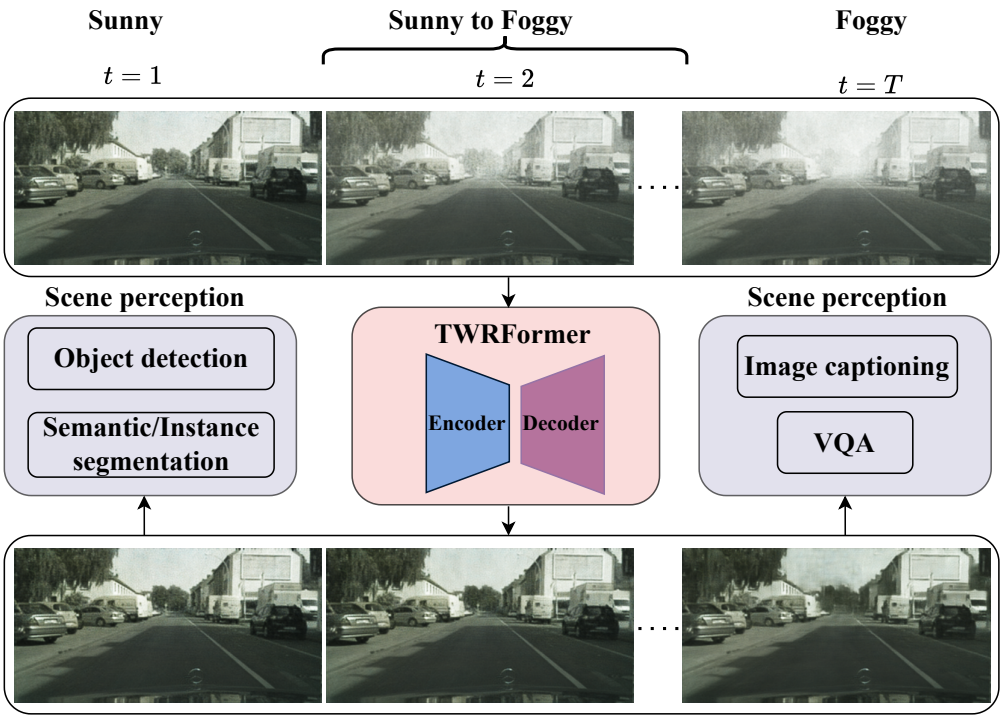


Figure 1: Overview of TransWeatherNet: Transitional weather images are first restored by TWRFormer and subsequently processed by detection, segmentation, and VLM backbones to improve autonomous driving perception under dynamically changing weather conditions.

pedestrians, and anomalous events, thereby mitigating operational risks and bolstering public safety. *(iii)* Beyond immediate perceptual gains, high-quality restored images serve as an indispensable foundation for a broad spectrum of downstream tasks, including object detection, semantic segmentation, and the rapidly advancing paradigm of vision-language models (VLMs), whose performance is sensitive to input image quality (Lin et al., 2025).

Need for Transitional Weather. Transitional weather refers to intermediate states that arise when one weather condition gradually evolves into another, such as fog turning into rain or rain clearing into sunshine, often with varying levels of intensity (Sun et al., 2022; Kong et al., 2023; Madhavi et al., 2026a). These dynamic shifts are highly relevant to autonomous driving perception (ADP), as they introduce abrupt, spatially varying degradations, such as inconsistent illumination and partial occlusions, which together complicate both weather restoration and downstream perception tasks (Liu et al., 2024; Kondapally et al., 2026). Despite their importance, most existing datasets (Caesar et al., 2019; Zheng et al., 2026; Wang et al., 2025) focus primarily on clear or extreme weather conditions, largely overlooking the transitional scenarios encountered in real-world driving (Kondapally et al., 2024a; Madhavi et al., 2026b; Madhavi & Mohan, 2024). To address this gap, Kondapally et al. (2026) introduced AIWD16 dataset, which captures sixteen diverse transitional weather scenarios including cloudy to snowy (CSn), snowy to rainy (SnR), cloudy to foggy (CF), foggy to rainy (FR), sunny to rainy (SR), cloudy to rainy (CR), snowy to foggy (SnF), sunny to foggy (SF), as well as their reverse transitions. Each sequence of length $T = 10$ frames, with 293 transitions per state, results in a total of 46,880 images. This comprehensive coverage enables AIWD16 to capture the weather dynamics, providing a valuable benchmark for robust perception under transitional conditions. To the best of our knowledge, it is the first dataset specifically designed to support transitional weather perception for real-time ADP.

Impact of Transitional Weather on Scene Perception. Downstream perception tasks, such as road element detection, segmentation, captioning, and VLM, remain highly challenging in real-world driving scenarios due to dynamic traffic patterns, complex road geometries, and frequent external noise or occlusions

(Hwang et al., 2025; Shen et al., 2025; Madhavi et al., 2024). VLMs have recently demonstrated strong potential for interpretable scene understanding in AV technology, supporting multimodal tasks like image captioning ("A pedestrian is crossing the street") and VQA ("How many objects are on the road?") (Xing et al., 2025; Zhou et al., 2026). However, the robustness of these perception pipelines degrades markedly under adverse or transitional weather, where shifting conditions (e.g., fog evolving into rain) introduce rapid illumination changes, increased occlusion, and non-uniform visibility, all of which amplify domain shifts and obscure critical agents or boundaries (Galshetwar et al., 2026; Pan et al., 2025). Conventional models (Kong et al., 2025c; Xu et al., 2025), which are primarily trained under static or clear conditions, exhibit significant performance declines in such environments, underscoring the need for weather-aware frameworks that can both generate and restore scenes under transitional weather settings (Lu et al., 2025a; Shaffee & Reda, 2026).

Limitations of Existing Image Restoration Methods. Table 1 summarizes existing adverse weather image restoration methods (Kulkarni et al., 2026; Mandal et al., 2026; Kong et al., 2025a), while a detailed discussion of related work is provided in Section 2. Notable advances have been made, most approaches target either single-task (Kong et al., 2025b; Lin et al., 2024) or discrete multi-weather restoration (Sun et al., 2024; Luo et al., 2025a), handling degradations like rain, fog, or snow independently, thus failing to model the continuous transitions common in real-world driving. Recent multi-task methods (Hu et al., 2025; Zhou et al., 2025) improve generality but often incur high parameter costs due to multi-encoder branches or heavy attention, limiting scalability. Further, the prevalent focus on static images overlooks crucial temporal dependencies, resulting in inconsistent restoration across frames during transitional weather conditions. Further, existing methods (Luo et al., 2025a; Rajagopalan et al., 2025) optimize only low-level metrics (e.g., *PSNR*, *SSIM*), neglecting their effect on downstream perception tasks, so visually pleasing results may still lack cues necessary for semantic understanding. Finally, the scarcity of realistic transitional weather datasets hinders model development and evaluation under dynamic, mixed conditions.

To address these limitations, we propose TransWeatherNet, a modular two-stage framework for restoring transitional weather-degraded scenes and enhancing autonomous driving perception under dynamically evolving weather conditions. First, TWRFormer, a transformer-based restoration network with a lightweight depthwise-separable projection head, restores visual clarity and structural consistency while leveraging perceptual loss to preserve semantic fidelity. Second, downstream object detection, segmentation, and vision-language models are trained on weather-restored images produced by TWRFormer using the AIWD16 dataset and further fine-tuned on both seen-domain benchmarks (Foggy Cityscapes, Rainy Cityscapes, BDD100K) and the unseen-domain benchmark ACDC to evaluate generalization robustness. Figure 1 illustrates the overall framework. Unlike prior restoration methods that optimize pixel fidelity under static corruption assumptions, TransWeatherNet explicitly models transitional weather as a non-stationary degradation process and incorporates degradation-aware attention conditioning, perceptual feature supervision, and downstream semantic optimization to achieve weather-adaptive, perception-oriented restoration, enabling direct integration into large-scale autonomous driving pipelines. The main contributions are:

1. **Comprehensive framework for transitional weather perception.** We propose TransWeatherNet, a modular framework comprising two complementary stages, degradation-aware image restoration and downstream perception evaluation, that collectively enable systematic analysis and improvement of perceptual robustness under continuously evolving environmental conditions.

2. **Transition-aware restoration network.** We introduce TWRFormer, a transformer-based architecture that restores images degraded by transitional weather patterns (e.g., foggy to rainy, snowy to cloudy). TWRFormer combines an adaptive transformer encoder for long-range degradation modeling with a weather-aware decoder conditioned on learned weather embeddings, and employs a lightweight depthwise-separable projection head for efficient reconstruction. Perceptual loss further enforces structural and semantic consistency.

3. **Comprehensive cross-domain evaluation.** Extensive experiments on AIWD16 (Kondapally et al., 2026) and four public benchmarks (Foggy Cityscapes (Sakaridis et al., 2018), RainCityscapes (Hu et al., 2019), BDD100K (Yu et al., 2020), and ACDC (Sakaridis et al., 2021)) demonstrate that TWRFormer achieves state-of-the-art restoration quality while significantly improving downstream detection and seg-

mentation accuracy, including on the unseen-domain ACDC benchmark, validating cross-domain generalization.

Paper outline. Section 2 reviews prior works on weather restoration, scene perception under adverse weather, and weather synthesis in AVT. Section 3 discusses essential background information of the proposed method. Section 4 illustrates the details of our proposed method, which comprises two stages, weather restoration, and downstream perception. Section 5 discusses Software and Machine Setup. Section 6 presents a comprehensive assessment of our proposed method for weather restoration and downstream perception. Section 7 provides an ablation study of our work. Section 8 provides limitations of our work and future research directions. Finally, Section 9 summarizes key research findings.

2 Related Work

2.1 Single-weather Restoration

A long-standing line of research in image restoration for adverse weather conditions has focused on discrete, well-characterized degradations such as haze, rain, snow, fog, and raindrops. For example, Rong et al. (2025) proposed STRRNet, a semantics-guided two-stage network for raindrop removal under both day and night conditions. Their framework integrates a semantic guidance module and a background restoration module, applying median fusion of multiple frames to generate pseudo labels that robustly handle dense raindrops and blurred backgrounds. Likewise, Lai et al. (2025) introduced SnowMaster, which constructs a large real-world snowy-image dataset (RealSnow10K) and leverages a feedback based mechanism with multimodal large language models to evaluate and filter pseudo-labels, enabling semi-supervised desnowing that generalizes better to real snow scenes.

These approaches often assume paired clean and degraded data and train a dedicated network for each degradation type. While highly effective within their specific domains, they lack generalization across diverse or evolving weather states. Earlier works on rain or haze removal, for instance, relied on task-specific architectures that do not extend naturally when multiple degradations coexist or transition over time. Such limitations make them less suitable for autonomous driving scenarios, where environmental conditions can change dynamically and involve overlapping degradations.

2.2 All-in-one Weather Restoration

To overcome the limitations of single-weather models, recent research has shifted toward unified restoration frameworks that handle diverse degradations within a single network (Rajagopalan et al., 2025).

Language and prior-guided approaches. Yang et al. (2024) leverage pre-trained vision-language models to generate degradation priors and dynamically activate expert sub-modules via a Mixture-of-Experts design. Luo et al. (2025b) similarly encode degradation types on standardized visual templates to guide a diffusion-based restoration network. Chen et al. (2025b) couple a generative diffusion prior with a task-feature adaptation module to simultaneously enhance perceptual quality and downstream task relevance.

Architectural innovations. Tian et al. (2025a) address internal task interference through channel-wise and attention-wise perturbations to align feature spaces across degradation types. He et al. (2025) introduce a scale-adaptive deformable transformer with scale-enhanced deformable convolutions for structured artifact removal. Gu et al. (2025) replace state-space-model blocks with linear attention modules to achieve global receptive fields at reduced computational cost. Liu et al. (2025) extend unified restoration to ultra-high-definition images by decomposing the task into frequency bands and learning compact latent representations through a VAE.

Flow-based methods. Qin et al. (2025) model degradation as a deterministic invertible flow rather than a stochastic diffusion, enabling efficient restoration in only a few sampling steps. These unified frameworks collectively demonstrate that a single network can learn to restore multiple degradation types. However, they continue to model each degradation as a distinct, static condition and thus lack mechanisms to capture the smooth transitions and compound effects inherent in transitional weather states.

Table 1: Comparison of existing restoration methods with the proposed method. R→Rainy, C→Cloudy, S→Snowy, F→Foggy, Su→Sunny, H→Haze, , L→Low light. Also, ✓→Present, ✗→Not present, '-'→Not applicable, TC→Transition classes, Real data→transitions derived from real-world data.

Method	Method Type	AV Scenarios	Dataset	Max Weather Conditions	No. of Samples	Backbone	Downstream Evaluation			Transition, TC Real data
							OD	Seg	VLM	
UniRestore (Chen et al., 2025a)	Multi-task	✓	Rain100L, DIV2K	S, R, H	2000	VAE	✓	✓	✗	✗, -, -
PromptIR (Potlapalli et al., 2023)	Multi-task	✗	BSD400	R, H	400	Transformer	✗	✗	✗	✗, -, -
SADT (He et al., 2025)	Multi-task	✗	SPAD	R	638, 492	Depthwise convolution	✗	✗	✗	✗, -, -
DPIR (Kong et al., 2025b)	Task-specific	✗	DIV2K	L	2000	SD3	✗	✗	✗	✗, -, -
JarvisIR (Lin et al., 2025)	Multi-task	✓	CleanBench	R, L, H	150,000	Llava-Llama3-8b	✗	✗	✗	✗, -, -
ILAWR (Lu et al., 2025b)	Multi-task	✓	RESIDE, Rain100H	H, R, S	5000	Convolution	✗	✗	✗	✗, -, -
Defusion (Luo et al., 2025a)	Multi-task	✓	Rain1400	R, D	5000	U-Net	✗	✗	✗	✗, -, -
GenDeg (Rajagopalan et al., 2025)	Multi-task	✓	GenDS	S, R, H, L	550,000	CLIP	✗	✗	✗	✗, -, -
DFP (Tian et al., 2025a)	Multi-task	✗	BSD400	R, B, H	400	Transformer	✗	✗	✗	✗, -, -
TransWeather (Valanarasu et al., 2022)	Multi-task	✓	S100K, RainDrop	F, R, S	100,000	Transformer	✗	✗	✗	✗, -, -
HistoFormer (Sun et al., 2024)	Multi-task	✓	Snow100K, RainDrop	R, S	100,000	Transformer	✗	✗	✗	✗, -, -
DiffBIR (Lin et al., 2024)	Task-specific	✗	laion2b-en	L	230000	Transformer	✗	✗	✗	✗, -, -
TWRFormer (Ours)	**Multi-task**	✓	**AIWD16 (Kondapally et al., 2026)**	**R, S, F, C, Su**	**46880**	**Transformer**	✓	✓	✓	✓, **16**, ✓
			FCS (Sakaridis et al., 2018) RCS (Hu et al., 2019), BDD100K (Yu et al., 2020)	**C, R, F**	**5000**					✗, -, -

2.3 Scene Perception Under Adverse Weather

Adverse weather conditions significantly degrade downstream perception by introducing visibility loss, appearance distortions, and domain shifts. Existing research has primarily addressed this challenge through domain adaptation, condition-invariant representation learning, and more recently, perception under transitional weather.

Domain adaptation. A large body of work seeks to bridge the distribution gap between clear-weather and adverse-weather domains. Zhao et al. (2024) proposed a universal mixing operator to improve generalization across diverse weather conditions, while Pan et al. (2025) disentangled weather-specific and weather-invariant features for robust adaptation. To reduce reliance on paired supervision, Li et al. (2023b) introduced a reference-free normal-to-adverse adaptation framework, and Wang et al. (2023) combined distribution-level and semantic-level adaptation in a backpropagation-free setting. More recently, Kondapally et al. (2024a) developed a multi-scale adaptation strategy specifically targeting transitional weather scenarios. Despite these advances, most adaptation methods assume predefined source and target domains. In practice, transitional weather introduces continuously evolving degradations and varying intensity levels, resulting in substantial intra-domain variability and non-stationary domain shifts that challenge conventional adaptation strategies.

Condition-invariant representations. Another line of research aims to learn weather-robust representations that remain stable across environmental conditions. Brüggemann et al. (2023) employed contrastive learning to improve semantic segmentation under varying weather and illumination conditions. Ding et al. (2023) proposed a hierarchical grouping transformer that captures multi-scale semantic structures for cross-domain segmentation. More recently, Zhang & Tan (2025) leveraged the complementary strengths of vision foundation models (VFMs) and vision-language models (VLMs) through a Mamba-based fusion architecture to achieve scalable multimodal alignment across adverse-weather domains.

Although these methods improve perception robustness, they generally operate directly on degraded inputs and focus on adapting perception models to weather-induced domain shifts. Consequently, their performance remains sensitive to severe visibility degradation and weather conditions not observed during training. Furthermore, most existing approaches assume static weather categories, overlooking transitional weather where degradations evolve continuously over time.

Transitional-weather perception. Recent studies have extended perception beyond static adverse weather by explicitly considering weather transitions, where environmental conditions evolve continuously over time. Prior work investigated scene recognition under changing weather conditions (Kondapally et al., 2024b) and spatio-temporal weather forecasting through TWFNet (Madhavi et al., 2026b). For downstream perception, Kondapally et al. (2024a) examined object detection under transitional weather, while CaRS (Madhavi et al., 2026a) introduced a causal intervention framework for semantic segmentation. More recently, Kondapally et al. (2026) explored vision-language assistance under evolving weather conditions, and

TransWardX (Madhavi et al., 2024) studied robustness through explainable black-box attacks on object detectors.

Although these methods improve robustness and perception under challenging environments, they generally operate directly on degraded observations and focus on adaptation, forecasting, recognition, segmentation, multimodal reasoning, or robustness analysis. Consequently, the influence of restoration under continuously evolving weather transitions remains insufficiently explored. Furthermore, most existing restoration methods are designed for static weather categories and are not explicitly optimized for transitional weather dynamics. In contrast, our framework introduces a dedicated restoration stage that models evolving weather degradations before downstream perception, enabling consistent improvements across detection, segmentation, and vision-language tasks.

2.4 Adverse Weather Datasets

Data scarcity remains a key bottleneck for training robust perception models. Several works have proposed synthetic weather generation to address this gap. Marathe et al. (2023) developed the WEDGE dataset, comprising images across sixteen extreme weather conditions. Zhu et al. (2023) introduced the SPA+ and RealSnow datasets, designed for deraining and desnowing tasks, respectively. Rothmeier & Huber (2021) and Mușat et al. (2021) employed GAN and CycleGAN frameworks to simulate fog, rain, and snow, while Hu et al. (2021) synthesized rain and introduced RainCityscapes. Wang et al. (2021a) proposed VRGNet for dynamic rain-streak generation, and Li et al. (2021) introduced WeaGAN for multi-domain weather translation. More recent efforts have moved toward spatiotemporally consistent synthesis, Li et al. (2025b) present DrivingDiffusion, a diffusion-based framework for generating realistic multi-view driving videos, while El-Shair et al. (2024) contribute a large-scale stereo dataset capturing diverse real-world conditions.

Moreover, Sun et al. (2022) introduced the SHIFT dataset, providing synthetic driving scenes with both discrete and continuous weather transitions for domain adaptation research. However, SHIFT primarily focuses on simulator-generated environments and was not designed for unified evaluation across multiple perception tasks under transitional weather. To facilitate such studies, Kondapally et al. (2024b) introduced AIWD6, a benchmark containing six transitional weather states. Building upon this effort, Kondapally et al. (2026) proposed AIWD16, comprising sixteen transitional weather states generated through conditional latent-space interpolation and accompanied by dense annotations supporting detection, segmentation, captioning, and vision-language evaluation.

Summary. Existing research on adverse-weather autonomous driving has progressed along three largely independent directions. Single-weather restoration methods achieve strong performance for specific degradations such as rain, snow, or haze, but require degradation-specific models and exhibit limited generalization across weather conditions. More recent all-in-one restoration frameworks unify multiple degradation types within a single architecture; however, they generally treat weather effects as discrete and static categories, without explicitly modeling the continuous evolution of weather transitions. Concurrently, perception-oriented studies have investigated domain adaptation, condition-invariant representations, forecasting, recognition, segmentation, and multimodal reasoning under adverse and transitional weather, while recent datasets have enabled systematic evaluation of these challenges.

Despite these advances, the interaction between restoration and downstream perception under continuously evolving weather conditions remains insufficiently explored. In particular, existing restoration approaches are not designed to capture the non-stationary characteristics of transitional weather, whereas perception methods typically operate directly on degraded observations. TransWeatherNet addresses this gap through a modular framework that combines transition-aware image restoration via TWRFormer with comprehensive downstream evaluation across object detection, semantic and instance segmentation, image captioning, and visual question answering, enabling robust perception under evolving weather conditions.

Table 2: Summary of adopted notations

Notation	Definition
$\mathcal{D}$	Training set
$\mathcal{M}$	Size of training set
x_i, y_i	Input image pair (Ex. x_i: Sunny image and y_i: Rainy image)
l_i	One-hot encoded labeled vector
Ψ	Variational parameters
z	Latent variable
u	Input features
Θ	Model parameters
μ	Mean of gaussian distribution
σ	Variance of gaussian distribution
β	Coefficient
$\mathcal{L}$	Loss
t	Transition sequence
T	Length of transition sequence
β_0 & β_1	Coefficients
ϵ	Error
$\hat{u}$	Reconstructed input
$\mathcal{V}$	Feature vectors obtained from CNN
$\mathcal{P}$	Processed feature vectors
α	Attention vector
$\mathcal{C}$	Combined context vector
h	Hidden state
I_{obs}	Degraded image
I_{gt}	Clean ground-truth image
I_{rest}	Restored image
$\theta(t)$	Transitional weather parameters
q_w	Weather query embeddings
F_Φ	Restoration network
$z_e^{(m)}$	Encoder feature at block m
$z_d^{(n)}$	Decoder feature at block n

3 Preliminaries

3.1 Transitional Weather Degradation

Image restoration under adverse weather has traditionally addressed discrete, well-defined degradation types such as haze, rain, snow, or fog, where each condition is modeled independently with task-specific or unified architectures. However, real-world weather does not conform to such categorical boundaries. Environmental conditions evolve continuously, for instance, from fog to rain or from overcast skies to snow, producing compound and time-varying degradations that span the spectrum between canonical weather states (Sun et al., 2022). These *transitional weather degradations* pose distinct challenges for image restoration:

- **Non-stationary degradation characteristics.** Unlike static weather, transitional conditions involve continuously shifting visibility, illumination, and contrast patterns. Restoration models trained on fixed degradation profiles struggle to generalize across these evolving distributions.

- **Compound and overlapping artifacts.** During transitions, multiple degradation types may coexist (e.g., residual fog combined with emerging rain streaks), producing artifact patterns that do not correspond to any single predefined category and are therefore poorly handled by existing restoration networks.

- **Absence of transitional training data.** Current restoration benchmarks (e.g., Foggy Cityscapes, RainCityscapes) capture only discrete weather endpoints. The lack of data representing intermediate transitional states limits the ability of restoration models to learn smooth, continuous degradation removal.

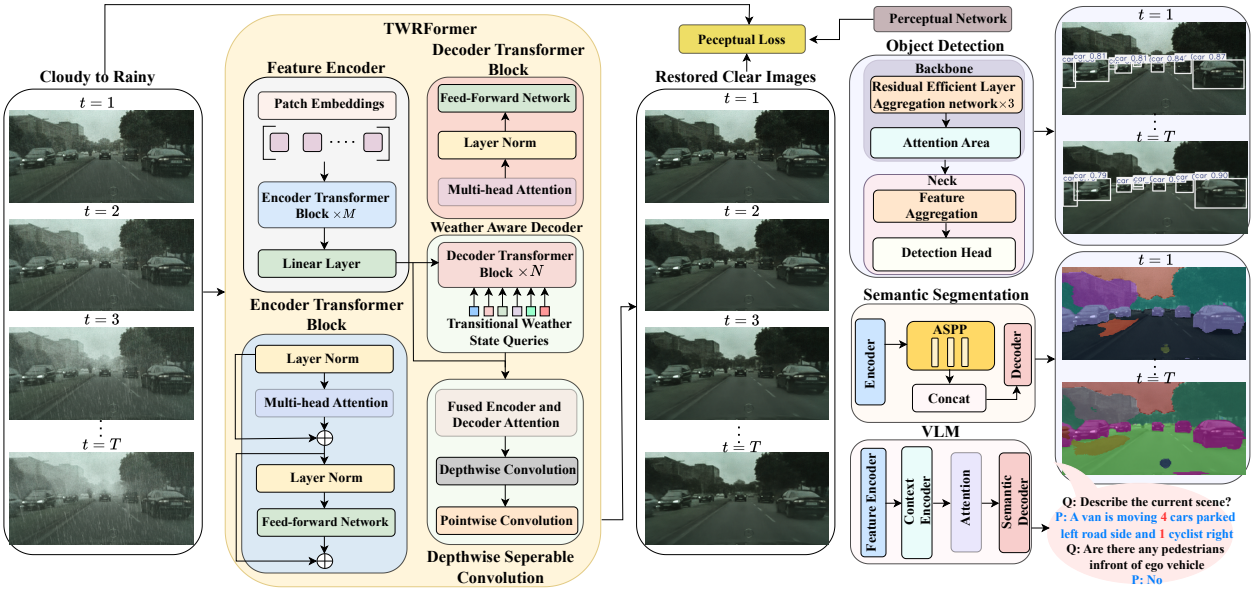


Figure 2: Architecture of TransWeatherNet. The transitional weather sequences (e.g., cloudy to rainy) from AIWD16 (Kondapally et al., 2026), which TWRFormer restores through a transformer encoder–decoder with depthwise separable projection head and perceptual-loss supervision. The restored images are then fed into detection, segmentation, and VLM backbones for robust perception under dynamic weather.

These limitations motivate the need for both a restoration architecture capable of adapting to evolving degradation profiles and a dataset that systematically captures transitional weather states for training and evaluation.

4 Proposed Method

Method overview. TransWeatherNet is a modular two-stage framework comprising transitional weather image restoration, and downstream perception evaluation. Each stage operates independently, enabling flexible training and systematic analysis of how weather synthesis quality propagates through restoration to high-level perception.

Stage 1: Image Restoration via TWRFormer. TWRFormer, an efficient encoder-decoder transformer architecture, employs a lightweight depthwise-separable projection head for image reconstruction and is additionally supervised with a perceptual loss computed through a frozen VGG-19 feature extractor to preserve high-level structural and semantic consistency.

Stage 2: Perception Evaluation. Restored images are passed to downstream detection, segmentation, and vision-language model (VLM) backbones to quantify the impact of restoration on recognition accuracy and scene understanding under evolving weather conditions. This stage validates that TWRFormer's restoration yields measurable gains in perception-critical tasks beyond pixel-level fidelity. An overview of the full pipeline is illustrated in Figure 2.

4.1 Transitional Weather Image Restoration

4.1.1 Problem Formulation

Let $I_{\text{gt}} \in \mathbb{R}^{H\times W\times 3}$ denote a clean image captured under nominal conditions. In transitional weather, the observed degraded image $I_{\text{obs}} \in \mathbb{R}^{H\times W\times 3}$ is modeled as:

$$I_{\text{obs}} = \mathcal{D}\big(I_{\text{gt}},\, \theta(t)\big) + \eta, \tag{1}$$

where $\mathcal{D}(\cdot)$ is a spatially variant, non-linear degradation function, $\theta(t)$ denotes time-dependent degradation parameters encoding the evolving atmospheric state at time t, and η represents additive sensor noise. Unlike conventional static weather models that assume fixed atmospheric parameters (e.g., uniform scattering coefficients or constant rain-streak density), the temporal variance and spatial non-uniformity of $\theta(t)$ make the degradation function significantly more complex to invert.

The restoration task is to learn an inverse mapping $\mathcal{F}_\Phi : I_{\text{obs}} \mapsto I_{\text{rest}}$, parameterized by network weights Φ, such that:

$$I_{\text{rest}} = \mathcal{F}_\Phi(I_{\text{obs}}) \approx I_{\text{gt}}. \tag{2}$$

This constitutes an ill-posed inverse problem, as the non-stationary nature of $\theta(t)$ introduces additional ambiguity beyond that of static degradation models: the same spatial region may exhibit different degradation characteristics at different time steps, and multiple plausible clean reconstructions may be consistent with a single observation. To regularize this ambiguity, we incorporate perceptual priors and semantic supervision into the optimization objective, as detailed in Section 4.1.2.

4.1.2 Transitional Weather Restoration Transformer (TWRFormer)

Standard transformer-based restoration models apply globally shared attention distributions, implicitly assuming spatially stationary degradation. However, transitional weather produces non-stationary degradation whose statistics evolve across both spatial regions and time. To enable degradation-adaptive restoration, TWRFormer introduces learnable weather query embeddings that condition the decoder's attention on the current atmospheric transition state.

Feature encoder. Given I_{obs}, the encoder extracts hierarchical features through a stack of transformer blocks. At the m-th block, features are computed via multi-head self-attention (MHSA) and a feed-forward network (FFN):

$$\hat{z}_e^{(m)} = \text{LN}\big(z_e^{(m-1)} + \text{MHSA}(z_e^{(m-1)})\big), \tag{3}$$

$$z_e^{(m)} = \text{LN}\big(\hat{z}_e^{(m)} + \text{FFN}(\hat{z}_e^{(m)})\big), \tag{4}$$

where $\text{LN}(\cdot)$ denotes layer normalization. The FFN incorporates depthwise convolutions to inject local inductive biases essential for preserving fine-grained spatial structures (e.g., object boundaries, lane markings) that are critical for downstream perception.

Weather-aware decoder. The decoder introduces learnable transitional weather query embeddings $q_w \in \mathbb{R}^{M\times C}$, where M is the number of weather tokens and C is the feature dimension. These embeddings encode degradation-specific atmospheric priors and are jointly optimized with all network parameters during training. At the n-th decoder block, restoration is conditioned through weather-guided cross-attention:

$$\hat{z}_d^{(n)} = \text{LN}\big(z_d^{(n-1)} + \text{MHCA}(z_d^{(n-1)}, q_w)\big), \tag{5}$$

$$z_d^{(n)} = \text{LN}\big(\hat{z}_d^{(n)} + \text{FFN}(\hat{z}_d^{(n)})\big), \tag{6}$$

where $\text{MHCA}(\cdot, q_w)$ denotes multi-head cross-attention with q_w serving as the conditioning source. This formulation enables the decoder to dynamically modulate encoded features according to learned weather priors, providing an adaptive approximation to the inverse degradation process $\mathcal{D}^{-1}$.

Depthwise separable projection head. The fused encoder-decoder features f_{fused} are projected back to image space through a lightweight depthwise separable convolution (DSC) module:

$$I_{\text{rest}} = \tanh\big(\text{Conv}_{1\times1}(\text{DSC}(f_{\text{fused}}))\big). \tag{7}$$

Restoration objective. Standard pixel-wise losses often yield over-smoothed reconstructions that degrade downstream perception accuracy. To balance spatial fidelity with perceptual quality, we optimize a hybrid objective:

$$\mathcal{L}_{\text{rest}} = \lambda_1 \|I_{\text{rest}} - I_{\text{gt}}\|_1 + \lambda_2\, \mathcal{L}_{\text{perc}}, \tag{8}$$

where the ℓ_1 norm enforces pixel-level fidelity. The perceptual loss $\mathcal{L}_{\text{perc}}$ is computed using feature activations from a frozen, pre-trained VGG-19 network:

$$\mathcal{L}_{\text{perc}} = \sum_{l \in \mathcal{S}} w_l \left\| \phi_l(I_{\text{rest}}) - \phi_l(I_{\text{gt}}) \right\|_2^2, \tag{9}$$

where $\phi_l(\cdot)$ denotes the activation at layer l, $\mathcal{S} = \{\texttt{conv2_2}, \texttt{conv3_3}, \texttt{conv4_3}\}$, and w_l are uniform layer weights. This feature-space supervision encourages TWRFormer to produce restorations with sharp edges and realistic textures that preserve the semantic structures required by downstream detection and segmentation modules.

Computational complexity. For spatial token count N and feature dimension C, standard self-attention incurs $\mathcal{O}(N^2C)$ complexity. The weather-aware decoder introduces cross-attention over M learnable weather tokens ($M \ll N$), adding only$\mathcal{O}(MNC)$, a negligible overhead relative to the encoder's quadratic cost.

4.2 Downstream Perception Evaluation

The third stage of the framework evaluates how restoration quality translates into downstream perception performance. Restored images produced by TWRFormer are used as inputs to three independently trained perception modules, object detection, semantic segmentation, and vision-language reasoning to comprehensively assess scene understanding under transitional weather conditions. Each perception model is fine-tuned on restored images to ensure domain alignment with TWRFormer's output distribution. An overview is shown in Figure 2.

4.2.1 Object Detection

We employ YOLOv12 (Tian et al., 2025b) as the detection backbone, which combines a residual efficient layer aggregation network (R-ELAN) for multi-scale spatial representation with an attention-centric design for robust feature aggregation. Given a restored image I_{rest}, the detector produces a set of predictions:

$$\mathcal{B}(I_{\text{rest}}) = \{(b_k,\, s_k,\, c_k)\}_{k=1}^{K}, \tag{10}$$

where b_k, s_k, and c_k denote the predicted bounding box, confidence score, and category label, respectively. The detector is fine-tuned on restored images to align with the appearance characteristics of TWRFormer's outputs.

4.2.2 Semantic Segmentation

For dense scene parsing, we adopt DeepLabv3+ (Chen et al., 2018a), which employs Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context aggregation and a decoder with skip connections for boundary refinement. The model produces a per-pixel semantic prediction:

$$\mathcal{S}(I_{\text{rest}}) = \{p_i \mid i \in \Omega\}, \tag{11}$$

where p_i is the predicted class distribution for pixel i over the image domain Ω. Restored inputs reduce boundary ambiguity caused by weather-induced artifacts, improving delineation of overlapping scene elements such as vehicles, pedestrians, and infrastructure.

4.2.3 Vision-Language Reasoning

To evaluate high-level semantic understanding, we design a lightweight multimodal model for image captioning and visual question answering (VQA). The architecture comprises four components:

- *Weather-Aware Visual Encoder (WVE):* A MobileNetV2 backbone extracts compact visual features $v = \Phi(I_{\text{rest}}) \in \mathbb{R}^{N \times d}$ from the restored image I_{rest}, where $\Phi(\cdot)$ denotes the visual encoder, N is the number of spatial feature tokens, and d is the feature dimension.

- *Text Context Encoder (TCE):* Stacked LSTMs encode the input text (caption prefix or question q) into a context-aware linguistic representation $t = \Psi(q) \in \mathbb{R}^{L \times d}$, where $\Psi(\cdot)$ denotes the text encoder and L is the sequence length.
- *Cross-Modal Attention Fusion (CAFM):* Visual and textual representations are fused via scaled dot-product attention. Attention weights are computed as: $\alpha_i = \frac{\exp(v_i^\top t_j/\sqrt{d})}{\sum_{k=1}^{N} \exp(v_k^\top t_j/\sqrt{d})}$, where α_i denotes the attention coefficient for the i-th visual feature $\mathcal{V}_i$. The resulting context vector aggregates visually salient information:

$$\mathbf{v}_{\text{context}} = \sum_{i=1}^{N} \alpha_i \, \mathcal{V}_i. \tag{12}$$

  In matrix form, this is equivalently expressed as:

$$A(v,t) = \text{Softmax}\left(\frac{v\,t^\top}{\sqrt{d}}\right), \qquad z = A(v,t)\,v, \tag{13}$$

  where z is the fused multimodal representation capturing weather-adaptive, query-relevant visual information.
- *Semantic Decoder (SD):* An LSTM decoder generates captions or answers conditioned on z. Given a restored image I_{rest} and query q, the model predicts:

$$a = \mathcal{M}\big(\Phi(I_{\text{rest}}),\, \Psi(q)\big), \tag{14}$$

  where $\mathcal{M}(\cdot)$ denotes the full multimodal inference pipeline (CAFM + SD). The model supports both captioning (e.g., "A van is moving beside parked cars") and VQA (e.g., "Is a pedestrian in front of the vehicle?").

This lightweight design enables efficient multimodal reasoning suitable for resource-constrained deployment while providing a diagnostic tool for evaluating whether restoration preserves the semantic content needed for language-grounded scene understanding.

4.2.4 Training Procedure

The modular stages of TransWeatherNet are trained independently in a sequential pipeline: TWRFormer is trained on transitional weather degraded images using the restoration objective $\mathcal{L}_{\text{rest}}$ (Eq. 8). Further, each downstream perception model, YOLOv12 (detection), DeepLabv3+ (segmentation), and the multi-modal VQA model, is independently fine-tuned on the restored images produced by TWRFormer, using its respective task-specific loss ($\mathcal{L}_{\text{det}}$, $\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{vlm}}$).

This modular training strategy offers two practical advantages: (*i*) each component can be independently updated or replaced without retraining the full pipeline, and (*ii*) it avoids the gradient interference and hyperparameter sensitivity associated with multi-task joint optimization, enabling stable training across heterogeneous task objectives.

4.3 Unified Rationale of the Proposed Approach and Model Choices

The design of TransWeatherNet is guided by three principles: (*i*) transitional weather must be modeled as a continuous process, not a discrete label; (*ii*) restoration quality must be validated through downstream perception, not solely through pixel-level metrics; and (*iii*) each stage must remain independently trainable to support modularity, reproducibility, and component-level analysis.

Transformer-based restorer. Transitional weather produces spatially non-uniform degradations that combine global effects (illumination shifts, haze) with localized artifacts (raindrops, snow particles). CNNs, constrained by fixed receptive fields, struggle to jointly model both scales. Multi-encoder all-in-one architectures address multiple weather types but treat each as a discrete category, lacking adaptation to intermediate states. TWRFormer addresses this through self-attention for global context, depthwise convolutions for local structure, and weather query embeddings that condition restoration on the degradation profile without

Table 3: AIWD16 dataset (Kondapally et al., 2026) annotation details.

Task	No. of transitions	Seq. length (T)	No. of seq. annotated	No. of classes annotated	Class names	Annotation tool
Object detection	16	10	46880	8	car, person, bicycle, motorcycle bus, train, truck, and traffic light	LabelImg (Tzutalin, 2015)
Semantic				12	road, non-road	LabelImg (Tzutalin, 2015)
Instance					car, pedestrian, bicycle, bike bus, obstacle, rider, scooter tram, traffic light, van, truck	Roboflow (Dwyer et al., 2025)
Image captioning						Manual annotation
VQA			234400			

requiring explicit weather labels at inference. The depthwise-separable projection head and compact query set ($M \ll N$) maintain computational efficiency suitable for real-time deployment.

Perception backbones. The choice of downstream models reflects the practical constraints of autonomous driving, where perception must be both accurate and efficient. YOLOv12 provides a favorable accuracy-latency tradeoff for real-time detection. DeepLabv3+ offers mature multi-scale segmentation with strong boundary precision under dense degradations. For vision-language reasoning, we design a lightweight multi-modal architecture rather than adopting large foundation models (e.g., LLaVA, GPT-4V), as the latter are impractical for on-vehicle deployment due to memory and latency constraints. This choice enables direct evaluation of whether restoration improves language-grounded scene understanding within resource-feasible bounds.

5 Software and Machine Setup

All experiments are conducted using Python 3.8 with PyTorch and TensorFlow on an NVIDIA RTX A6000 GPU (48 GB). For image restoration, TWRFormer is trained for 100 epochs with a batch size of 32 and a learning rate of 2e-4. The model is supervised using a hybrid loss combining ℓ_1 reconstruction and perceptual loss extracted from a frozen VGG-19 network. All datasets are split into 70% training, 10% validation, and 20% testing. All baseline restoration models are trained under the same settings for fair comparison. For downstream perception, we train state-of-the-art backbones on the restored images. YOLOv12 is trained for 80 epochs with a batch size of 16 and a cosine-annealed learning rate of 1e-4 for object detection. For semantic segmentation, DeepLabv3+ is trained for 60 epochs with a learning rate of 6e-5 using AdamW. For both image captioning and VQA, our lightweight VLM is trained for 40 epochs with learning rates of 2e-4 for the visual encoder and 1e-3 for the LSTM-based decoder.

6 Experiments

6.1 Datasets and Implementation Details

Table 3 provides an overview of the AIWD16 (Kondapally et al., 2026), outlining the transitional weather states, annotations, object classes, and tools used. The AIWD16 dataset is constructed by generating transitional weather sequences using Stochastic Conditional Variational Autoencoder(SC-VAE). The SC-VAE operates on paired images captured under different weather conditions, producing temporally smooth transitions between weather domains. In total, 16 transitions are included, for example, cloudy to rainy (CR), sunny to foggy (SF), cloudy to snowy (CSn), snowy to rainy (SnR), foggy to rainy (FR), and vice versa. The full dataset occupies 7GB. Figures 3 and 4 illustrate example transitions. The generated sequences exhibit smooth variations in appearance across time (e.g., gradual accumulation of rain streaks and diffusion of fog), demonstrating that the model captures continuous, realistic transformations rather than abrupt or discrete changes. Table 4 summarizes the implementation details of TWRFormer, including backbone, training strategy, number of epochs, and key hyperparameters.

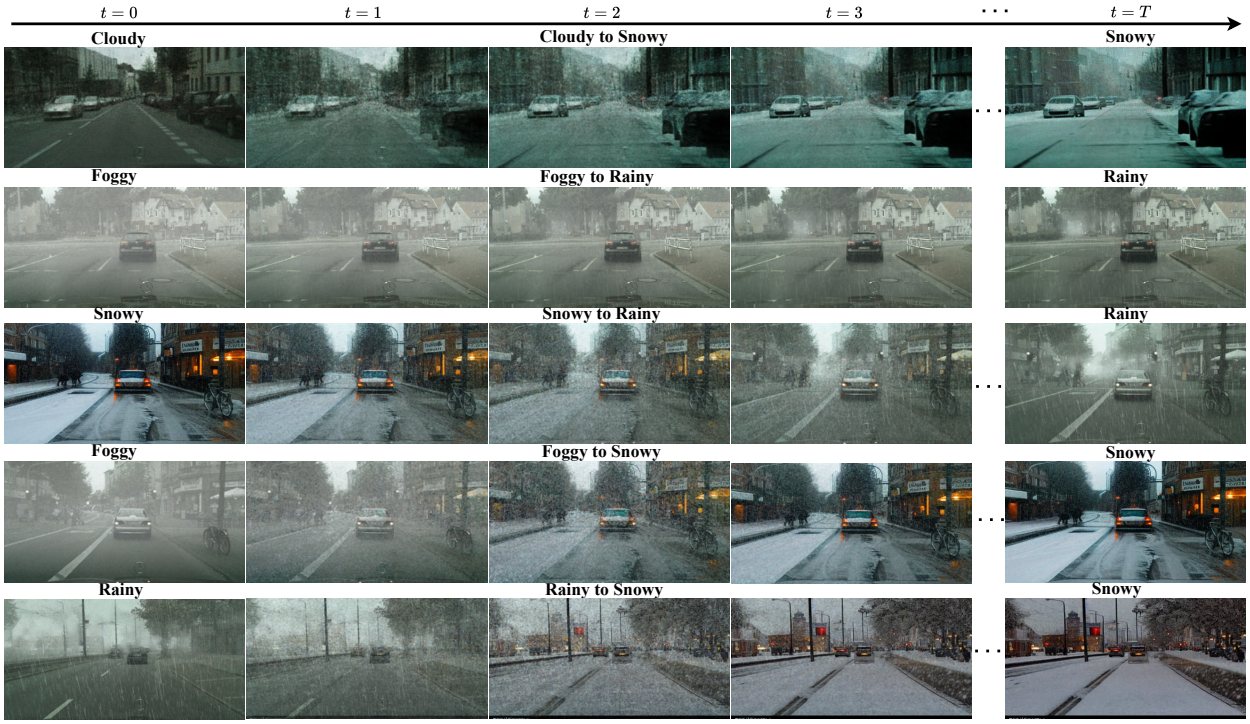


Figure 3: Transitional weather sequences (e.g., cloudy to snowy, foggy to rainy, snowy to rainy, foggy to snowy, and rainy to snowy) of the AIWD16 dataset.

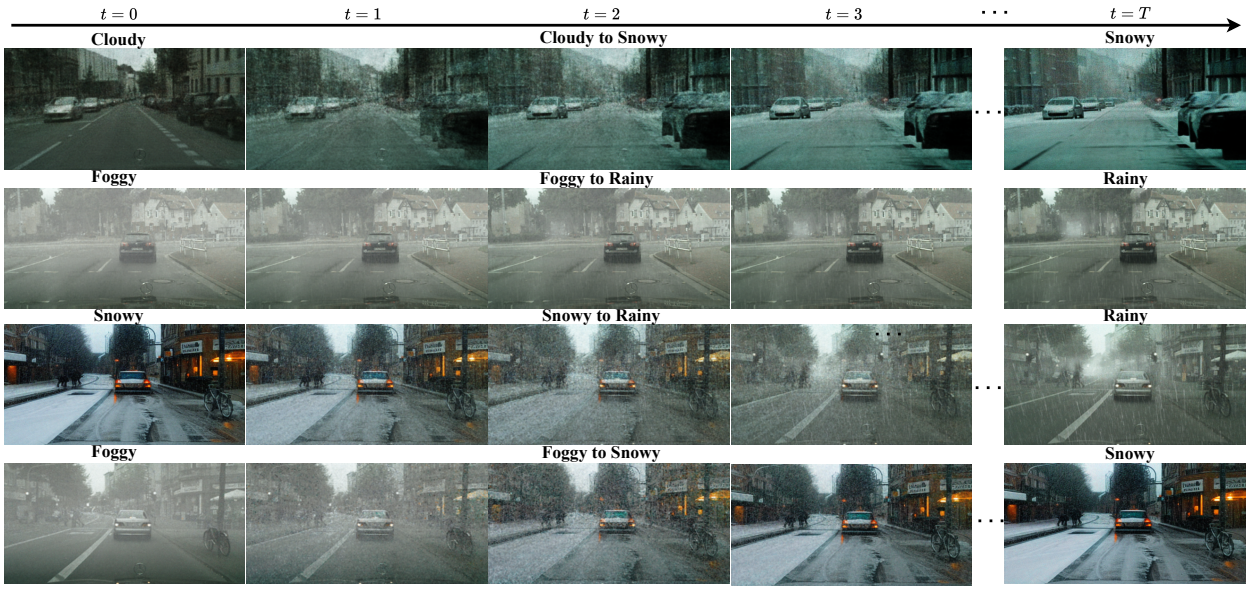


Figure 4: Transitional weather sequences (e.g., cloudy to snowy, foggy to rainy, snowy to rainy, foggy to snowy, and rainy to snowy) of the AIWD16 dataset.

Table 4: Implementation details of the proposed TransFormer approach. TS→Training speed, lr→Learning rate, op→Optimizer, l→Loss, b→Batch size, I→Input shape, CE→Cross Entropy, and AH→Attention Heads.

Backbone	**Epochs**	**Hyper parameters**	**TS (sec)**
Transformer with DSC	100	*lr:* 0.0002, *op:* Adam *l:* CE, *I:* 256 × 256, *b:* 32, AH:8	5460

Dataset annotations. The AIWD16 dataset (Kondapally et al., 2026) comprises five types of annotations designed to support a wide range of downstream perception tasks, including object detection, semantic segmentation, instance segmentation, image captioning, and visual question answering (VQA). For object detection, bounding-box annotations cover key autonomous driving categories such as car, person, bicycle, motorcycle, bus, train, truck, and traffic light. Semantic segmentation annotations were produced using LabelMe and provide pixel-level labels for 12 driving-related classes encompassing road users, static infrastructure, and traffic elements. Instance segmentation annotations include object-level masks for commonly encountered traffic participants and scene elements, including car, truck, bus, bicycle, pedestrian, rider, tram, van, scooter, and traffic light. In addition to visual annotations, the dataset includes manually curated text-based labels to facilitate image captioning and VQA tasks. Each image is accompanied by descriptive captions summarizing salient traffic elements, such as moving vehicles, pedestrians in the vicinity of the ego vehicle, and roadside parked objects. For VQA, five open-ended question–answer pairs are provided per image, covering scene understanding aspects such as object counts, vehicle categories, and pedestrian activities. Collectively, these multimodal annotations enable comprehensive evaluation of perception and scene understanding algorithms under transitional weather conditions.

6.2 Evaluation Metrics

For quantitative evaluation of image restoration, we employ the Peak Signal-to-Noise Ratio (*PSNR*) and the Structural Similarity Index Measure (*SSIM*) to assess reconstruction fidelity and perceptual quality. The performance of object detection is evaluated using the mean Average Precision (*mAP*), while semantic segmentation is measured using the mean Intersection over Union (*mIoU*). For vision–language tasks, including image captioning and VQA, we report BLEU and ROUGE scores to quantify linguistic relevance and semantic alignment between generated and ground-truth responses.

6.3 Comparison with Baselines

6.3.1 Image Restoration

To evaluate the effectiveness of the proposed TWRFormer, we compare it against task-specific restoration models (NAFNet (Chen et al., 2022a), Restormer (Zamir et al., 2022), SAPNet (Zheng et al., 2022)) and multi-task restoration approaches (TransWeather (Valanarasu et al., 2022), UniRestore (Chen et al., 2025a), DiffUIR (Zheng et al., 2024), Histoformer (Sun et al., 2024)). Table 5 reports quantitative results on AIWD16 and three public benchmarks. For AIWD16, all models are trained and evaluated under identical conditions. In contrast, for other datasets (Snow100K-L (Liu et al., 2018) and RainDrop (Qian et al., 2018)), models are fine-tuned on their respective training splits. We additionally evaluate performance on the unseen Test1 dataset (Li et al., 2019) to assess generalization, as shown in Table 6. Across all benchmarks, including unseen data, TWRFormer consistently surpasses both task-specific and multi-weather baselines in *PSNR* and *SSIM*, demonstrating strong performance with improvements of +2.01% in *PSNR* and +0.03% in *SSIM* for transitional weather restoration.

Further, Table 7 reports the image restoration performance of TWRFormer against baseline methods in terms of *PSNR* and *SSIM*, demonstrating that our method effectively handles all degradation types across a wide range of intensities. Figure 5 further illustrates the qualitative restoration results for several weather

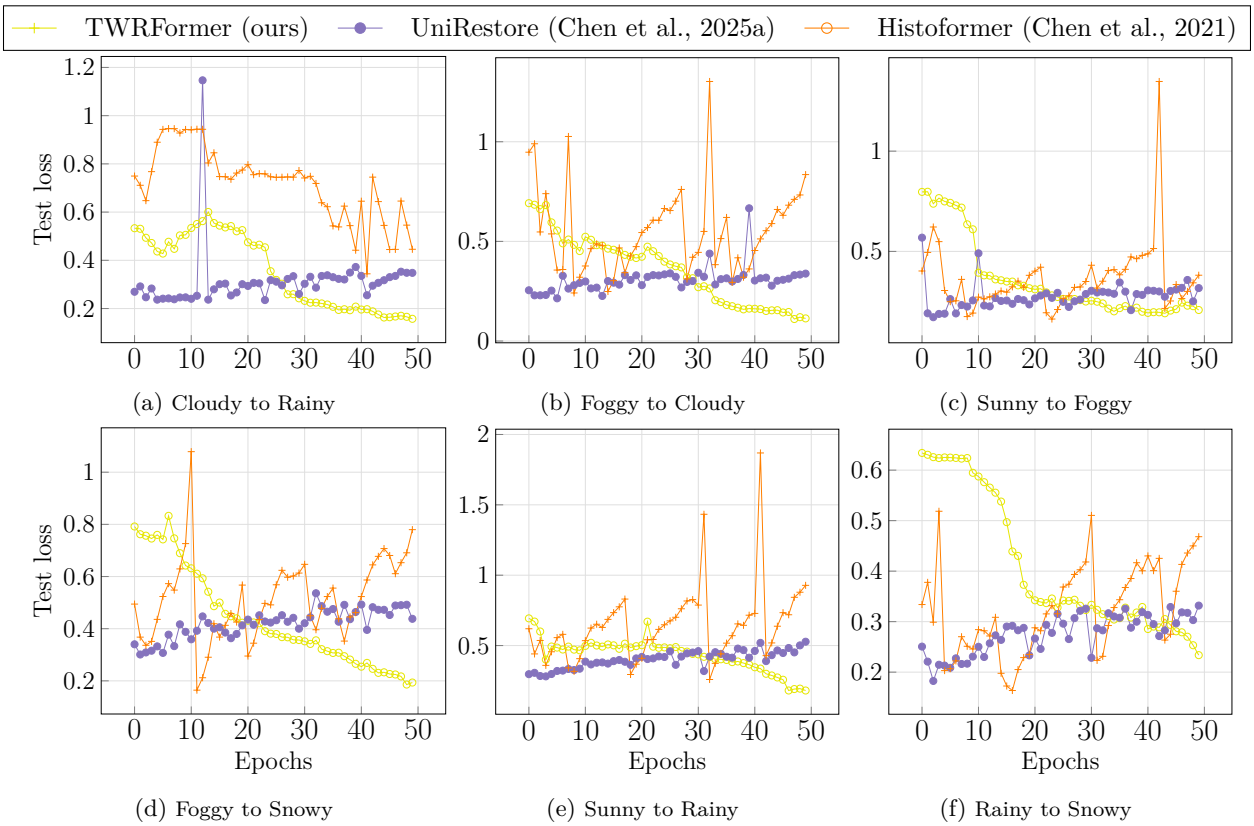
Table 5: Quantitative comparison on multi-weather restoration task in terms of *PSNR* and *SSIM*, where higher values indicate better performance. The table provides the analysis of task-specific and unified multi-weather models on the seen datasets. The best results are in **bold**.

Task type	Method	AIWD16 (Kondapally et al., 2026)		Method	Snow100K-L (Liu et al., 2018)		Method	RainDrop (Qian et al., 2018)	
		PSNR	SSIM		PSNR	SSIM		PSNR	SSIM
Task Specific	NAFNet (Chen et al., 2022a)	20.72	0.70	RESCAN (Li et al., 2018)	26.08	0.81	DuRN (Liu et al., 2019)	31.24	0.92
	Swin-IR (Liang et al., 2021)	19.91	0.78	DDMSNet (Zhang et al., 2021)	28.85	0.87	MAXIM (Tu et al., 2022)	31.87	0.93
	Restormer (Zamir et al., 2022)	20.77	0.77	NAFNet (Chen et al., 2022a)	30.06	0.90	IDT (Xiao et al., 2022)	31.87	0.93
	SAPNet (Zheng et al., 2022)	28.40	0.85	Restormer (Zamir et al., 2022)	30.36	0.90	Restormer (Zamir et al., 2022)	32.18	0.94
Multi Task	All-in-One (Li et al., 2020)	19.36	0.75	All-in-One (Li et al., 2020)	28.33	0.88	All-in-One (Li et al., 2020)	31.12	0.92
	TransWeather (Valanarasu et al., 2022)	22.36	0.81	TransWeather (Valanarasu et al., 2022)	29.31	0.89	TransWeather (Valanarasu et al., 2022)	30.17	0.91
	Chen et al. (Chen et al., 2022b)	23.49	0.80	Chen et al. (Chen et al., 2022b)	30.22	0.90	Chen et al. (Chen et al., 2022b)	31.81	0.93
	WGWSNet (Zhu et al., 2023)	22.99	0.80	WGWSNet (Zhu et al., 2023)	30.16	0.90	WGWSNet (Zhu et al., 2023)	29.38	0.88
	WeatherDiff (Özdenizci & Legenstein, 2023)	24.38	0.81	WeatherDiff (Özdenizci & Legenstein, 2023)	29.58	0.89	WeatherDiff (Özdenizci & Legenstein, 2023)	30.71	0.93
	Histoformer (Sun et al., 2024)	26.32	0.85	Histoformer (Sun et al., 2024)	28.44	0.90	Histoformer (Sun et al., 2024)	30.02	0.92
	DiffUIR (Zheng et al., 2024)	23.79	0.83	DiffUIR (Zheng et al., 2024)	20.74	0.87	DiffUIR (Zheng et al., 2024)	28.25	0.91
	UniRestore (Chen et al., 2025a)	27.81	0.86	UniRestore (Chen et al., 2025a)	29.16	0.86	UniRestore (Chen et al., 2025a)	32.07	0.90
	TWRFormer (Ours)	**29.82**	**0.89**	**TWRFormer (Ours)**	**33.39**	**0.92**	**TWRFormer (Ours)**	**33.85**	**0.93**

Table 6: Quantitative comparison on multi-weather restoration task in terms of *PSNR* and *SSIM*, where higher values indicate better performance. The table provides the analysis of task-specific and unified multi-weather models on the unseen dataset. The best results are in **bold**.

Task type	Method	Test1 (rain+fog) (Li et al., 2019)	
		PSNR	SSIM
Task Specific	pix2pix (Isola et al., 2017)	19.09	0.71
	RESCAN (Li et al., 2018)	15.91	0.61
	Swin-IR (Liang et al., 2021)	23.23	0.86
	MPRNet (Zamir et al., 2021)	21.9	0.84
Multi Task	All-in-One (Li et al., 2020)	24.71	0.87
	TransWeather (Valanarasu et al., 2022)	26.83	0.88
	Chen et al. (Chen et al., 2022b)	27.27	0.87
	WGWSNet (Zhu et al., 2023)	27.32	0.88
	WeatherDiff (Özdenizci & Legenstein, 2023)	28.64	0.89
	Histoformer (Sun et al., 2024)	27.91	0.88
	DiffUIR (Zheng et al., 2024)	23.58	0.82
	UniRestore (Chen et al., 2025a)	29.08	0.89
	TWRFormer (Ours)	**30.33**	**0.89**

Figure 5: Qualitative results on the AIWD16 dataset (Kondapally et al., 2026) for a sequence length T, where our proposed method effectively removes transitional weather patterns (cloudy, rainy, foggy, and sunny) across various intensity levels.



(a) Cloudy to Rainy (b) Foggy to Cloudy (c) Sunny to Foggy

(d) Foggy to Snowy (e) Sunny to Rainy (f) Rainy to Snowy

Figure 6: Comparison of image restoration loss of the TWRFormer with state-of-the-art models on the AIWD16 dataset (Kondapally et al., 2026).

Table 7: Quality of the restored images of the AIWD16 images (Kondapally et al., 2026) measured using *PSNR*, and *SSIM* metrics. Our restoration model is a unified architecture that can restore all types of weather conditions in different variations. **Bold** indicates the best values.

Dataset/ Metrics	CR	RC	SF	FS	SR	RS	CF	FC	CSn	SnC	SnR	RSn	SnF	FSn	FR	RF
PSNR	**34.72**	34.55	28.85	28.63	24.88	25.29	30.44	30.56	28.15	26.87	26.29	24.87	26.31	27.05	27.98	28.05
SSIM	0.92	**0.94**	0.91	0.89	0.79	0.73	0.91	0.90	0.86	0.83	0.82	0.85	0.84	0.84	0.88	0.89

transitions (e.g., Cloudy to Rainy, Sunny to Foggy, and Sunny to Rainy). In addition, Figure 6 presents the restoration loss comparisons between our approach and baseline models across different transition states in the AIWD16 dataset. The results indicate that TWRFormer consistently outperforms baseline models, delivering higher-quality restorations across diverse weather degradations and transition weather scenarios. This shows that the model generalizes well and effectively recovers clean images even under challenging, mixed-intensity conditions.

Table 8 presents a comparison of computational complexity between the proposed method and state-of-the-art baselines across five key indicators: model parameters (param), FLOPs, memory utilization, FPS, and inference speed (time), showing that the proposed approach requires lower computational cost than most existing methods. The improved computational efficiency is due to the use of a lightweight architecture with optimized operations, which reduces redundant computations while maintaining effective feature representation.

Table 8: Comparison of computational complexity, model size, and inference efficiency of the proposed TWRFormer method with state-of-the-art image restoration methods under transitional weather conditions.

Model	Type	No. of Param	FLOPs (G)	Memory (GB)	FPS	Infer Time (ms)
Transweather	Transformer	40M	80	4	15	60
Histoformer	Transformer	25M	40	2	15	65
DiffUIR	Diffusion	75M	370	8	2	479
UniRestore	Diffusion	150	980	11	1	990
TWRFormer (Ours)	**Transformer**	**22M**	**32**	**2**	**19**	**51**

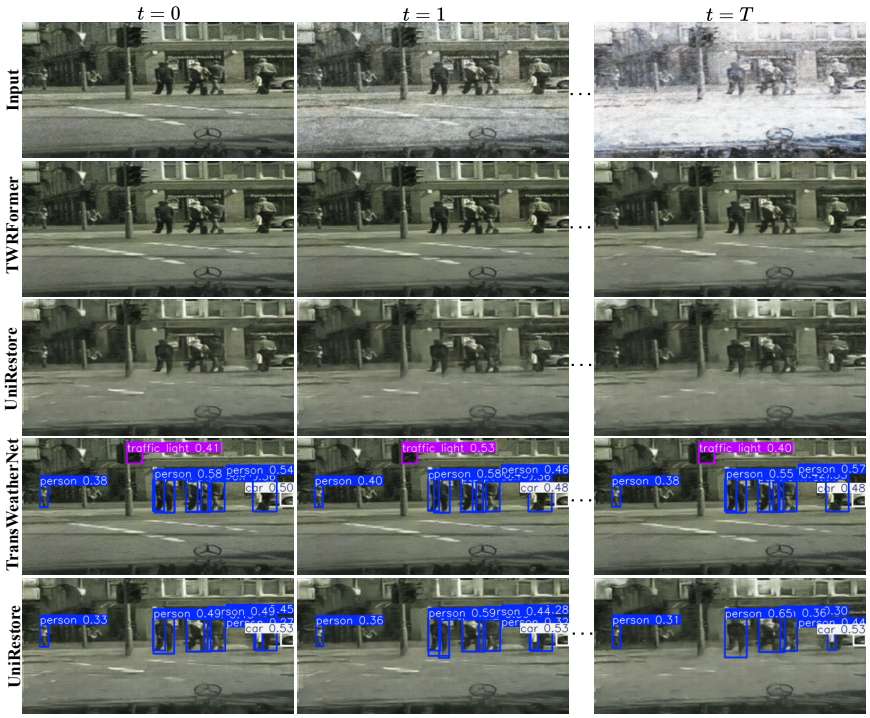


Figure 7: Qualitative results of transitional weather restoration and object detection on the AIWD16 dataset (Kondapally et al., 2026) for a sequence length T. TransWeatherNet effectively removes transitional weather artifacts (e.g., sunny to rainy) and enhances detection performance compared to baselines.

6.3.2 Downstream Perception Evaluation

Annotation statistics of AIWD16 (Kondapally et al., 2026) are summarized in Table 3. Using restored images significantly improves downstream perception by reducing weather-induced distortions, improving feature alignment, and preserving semantic consistency across dynamic scenes. Beyond AIWD16, we further evaluate performance on multiple seen and unseen benchmarks.

Object detection. Table 9 summarizes detection results under transitional weather. The first block reports baseline detectors (Wang et al., 2022; Carion et al., 2020) without restoration, the second highlights weather-specific detectors (Khindkar et al., 2022; Hnewa & Radha, 2023), and the final section compares detection performance using images restored by prior methods (Chen et al., 2025a; Zheng et al., 2024; Potlapalli et al., 2023) against our TransWeatherNet. In addition to AIWD16, we evaluate the approach on other benchmark datasets, Rainy Cityscapes (Hu et al., 2019), Foggy Cityscapes (Sakaridis et al., 2018), and BDD100K (Yu et al., 2020) with additional evaluation on the unseen ACDC dataset (Sakaridis et al., 2021) to assess generalization. Across all benchmarks, TransWeatherNet consistently achieves the highest mAP scores,

Table 9: Comparison with baselines on seen and unseen datasets for object detection. All results are presented in terms of mean Average Precision(mAP). FCS→Foggy Cityscapes, RCS→RainCityscapes, '-'→Not applicable.

Model	Seen dataset				Unseen dataset
	AIWD16 (Kondapally et al., 2026)	**RCS**	**FCS**	**BDD100K**	**ACDC**
YOLOv7 (Wang et al., 2022)	21.8	24.6	41.7	29.5	27.3
SSD (Liu et al., 2016)	23.1	19.7	32.3	25.4	24.2
DETR (Carion et al., 2020)	21.7	20.6	39.6	27.3	25.7
DAF (Chen et al., 2018b)	-	-	32.6	-	25.8
PDA (Hsu et al., 2020)	-	-	36.9	28.3	26.6
ICR-CCR (Xu et al., 2020)	-	-	37.4	28.9	27.2
MeGA (Vs et al., 2021)	-	-	41.8	-	27.4
DSS (Wang et al., 2021b)	-	-	40.9	-	28.6
ILLUME (Khindkar et al., 2022)	42.3	37.7	43.8	34.6	29.6
MS-DAYOLO (Hnewa & Radha, 2023)	41.8	35.6	38.3	32.2	29.9
DIP (Liu et al., 2022)	-	-	38.08	27.8	26.3
PromptIR (Potlapalli et al., 2023)	44.2	37.5	45.5	32.5	30.9
DiffUIR (Zheng et al., 2024)	42.5	38.7	44.9	33.7	31.2
UniRestore (Chen et al., 2025a)	46.4	37.9	45.5	34.7	32.6
TWRFormer (Ours)	**49.5**	**40.1**	**47.9**	**36.4**	**34.7**

demonstrating the significant benefit of high-quality restoration for reliable detection under transitional and mixed weather conditions.

Figure 7 shows qualitative comparisons on AIWD16, where TWRFormer more effectively removes complex, varying-intensity weather artifacts while preserving structural details compared to UniRestore (Chen et al., 2025a). These gains stem from TWRFormer's efficient feature fusion and robust modeling of long-range degradation patterns. These results further show that our restored images enable more accurate and complete detections than competing baselines. Some challenges remain for very small or distant objects, which we leave for future improvement.

Segmentation. We evaluate the impact of restoration on pixel-level scene understanding using semantic segmentation. Table 10 reports results across three groups: (i) state-of-the-art segmentation baselines (Chen et al., 2018a; 2021) applied directly to the generated dataset, (ii) weather-adaptation methods (Li et al., 2023b; Lee et al., 2022; Liu et al., 2022) that leverage domain adaptation or style transfer, and (iii) segmentation performance on images restored by existing approaches (Chen et al., 2025a; Zheng et al., 2024; Lin et al., 2024) compared with TransWeatherNet. Experiments on AIWD16, Rainy Cityscapes, Foggy Cityscapes, BDD100K, and the unseen ACDC dataset show that TransWeatherNet consistently yields the best mIoU, achieving a +3.5% gain on AIWD16. This demonstrates that higher-quality restoration substantially improves pixel-level understanding and preserves scene structure in transitional and mixed weather. Qualitative results in Figure 8 further highlight more complete and coherent segmentation maps produced by the proposed framework. Remaining challenges include fine-grained delineation of distant or occluded regions, which we plan to address through higher-resolution feature modeling in future work. Notably, unlike prior pipelines that treat weather adaptation, image restoration, and scene perception as isolated tasks, our modular framework coherently integrates all three stages, generation, restoration, and perception, enabling systematic evaluation of how restoration quality propagates to downstream task performance.

Vision language modeling. Table 11 compares TransWeatherNet with state-of-the-art VLMs on AIWD16 for captioning and VQA. Our method achieves +2.01% BLEU and +0.65% ROUGE-L improvements for captioning, and +0.8% BLEU and +0.93% ROUGE-L for VQA, while using significantly fewer parameters. Additional evaluations on DriveLM and the unseen NuScenes VQA dataset, as shown in Table 12, confirm strong generalization and robustness under transitional weather. The proposed lightweight VLM operates with only 12M parameters, 810M FLOPs, and 0.25 GB memory, representing reductions of 85% in parameters and 73% in memory compared to existing VLMs. This efficiency enables faster inference, higher FPS, and deployment on low-resource platforms without compromising accuracy. Figure 8 shows qualitative VQA

Table 10: Comparison with baselines on seen and unseen datasets for semantic segmentation. All results are presented in terms of mean Intersection over Union (mIoU). FCS→Foggy Cityscapes, RCS→RainCityscapes, '-'→Not applicable.

Model	Seen Dataset				Unseen Dataset
	AIWD16 (Kondapally et al., 2026)	**RCS** (Hu et al., 2019)	**FCS** (Sakaridis et al., 2018)	**BDD100K** (Yu et al., 2020)	**ACDC** (Sakaridis et al., 2021)
U-Net (Ronneberger et al., 2015)	41.52	37.45	55.35	32.54	25.66
TransUnet (Chen et al., 2021)	43.55	38.73	59.42	34.26	26.99
DeepLabV3+ (Chen et al., 2018a)	45.77	40.36	65.20	37.76	28.30
FIFO (Lee et al., 2022)	42.47	-	64.80	-	37.59
VBLC (Li et al., 2023b)	55.90	59.60	73.20	37.70	39.30
DIP (Liu et al., 2022)	54.21	56.19	67.81	46.59	38.19
DiffBIR (Lin et al., 2024)	51.43	53.42	66.29	45.25	36.28
URIE (Son et al., 2020)	54.45	54.81	65.93	42.32	37.90
NAFNet (Chen et al., 2022a)	56.90	57.45	66.90	46.25	39.22
DiffUIR (Zheng et al., 2024)	50.23	50.22	63.54	42.35	35.78
PromptIR (Potlapalli et al., 2023)	52.66	53.66	66.23	43.54	37.61
UniRestore (Chen et al., 2025a)	56.72	64.05	70.77	45.73	39.27
TWRFormer (Ours)	**60.40**	**65.75**	**74.66**	**47.91**	**40.11**

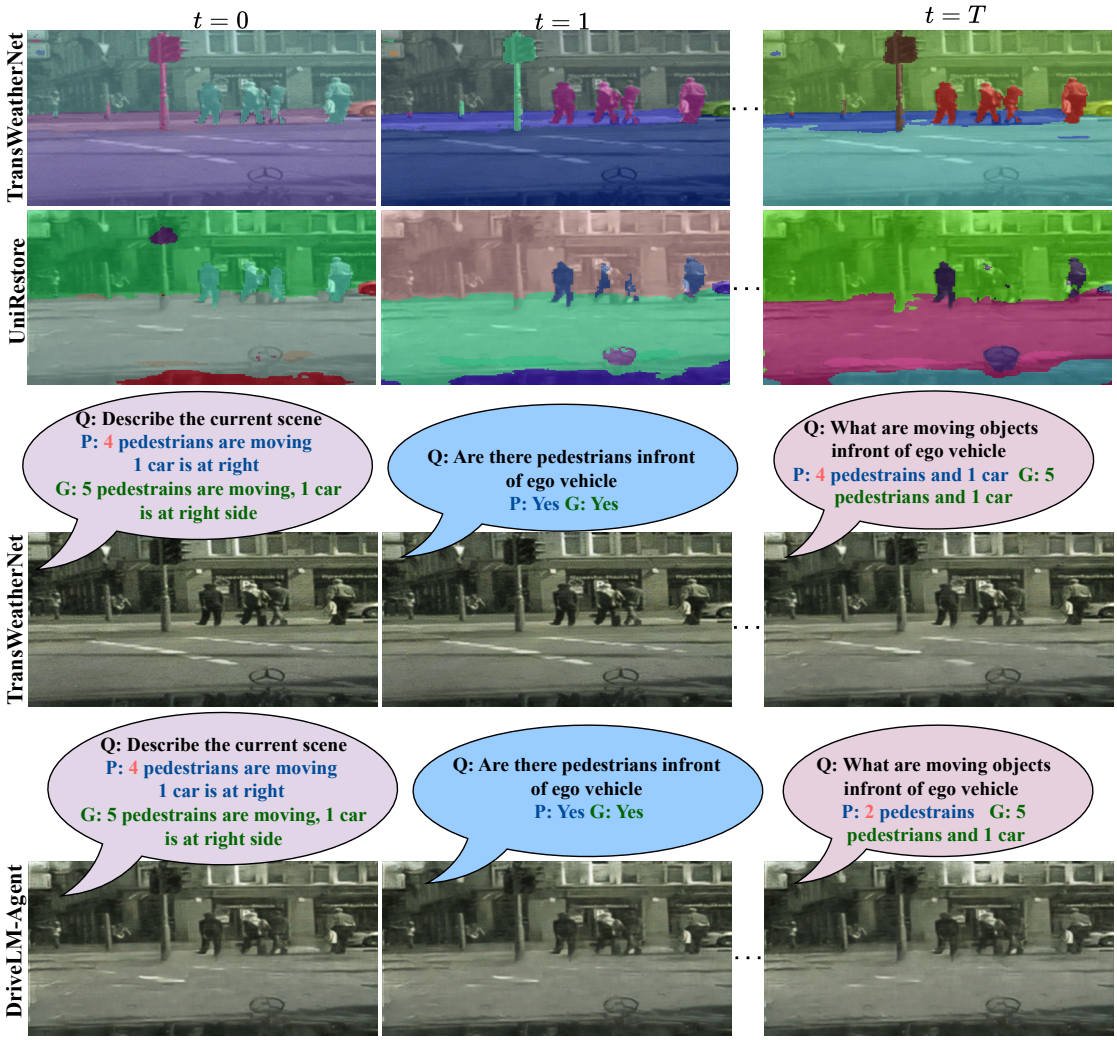


Figure 8: Qualitative results of semantic segmentation and visual question answering (VQA) on the AIWD16 dataset (Kondapally et al., 2026) for a sequence length T. TransWeatherNet significantly enhances segmentation quality and VQA accuracy compared to baselines.

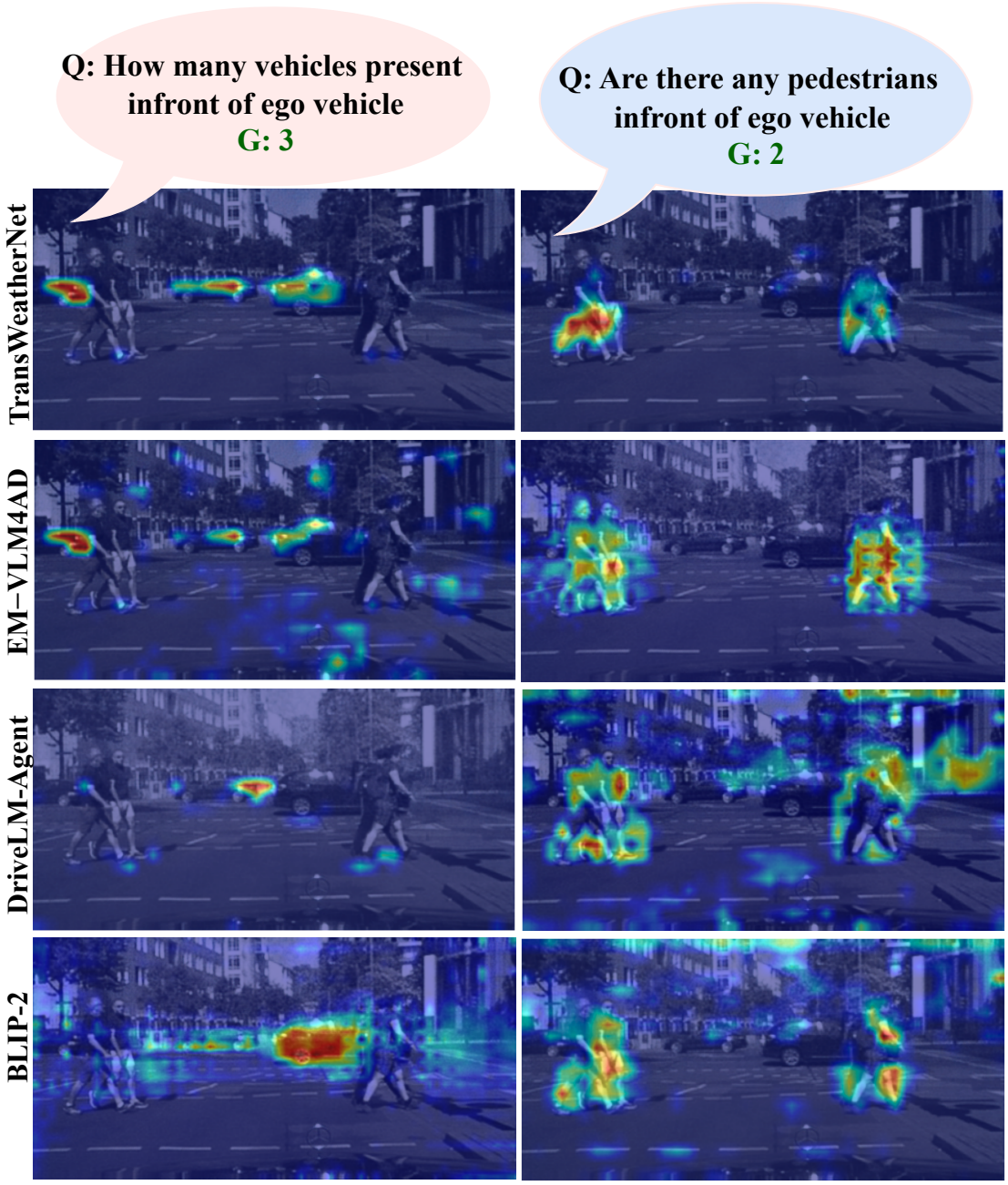


Figure 9: Visual comparison of attention heatmaps generated by TransWeatherNet and VLM baselines for the VQA task. The results show that our model provides more interpretable attention patterns and achieves better alignment with relevant scene regions, reflecting its performance gains over existing methods.

examples, where our approach yields more precise and context-aware responses. Furthermore, Figure 9 presents a visual comparison of attention heatmaps produced by TransWeatherNet and VLM baselines on the VQA task. The results show that our method yields more interpretable attention distributions and better aligns with scene regions relevant to the user queries, consistent with its performance improvements over existing approaches. Each example illustrates how the models allocate attention across road-scene objects pertinent to the given questions. Future work will target improvements in complex reasoning, multi-object interactions, and occlusion-heavy scenarios.

Table 11: Performance comparison between the proposed and state-of-the-art models for the image captioning and VQA tasks on the AIWD16 dataset (Kondapally et al., 2026). **Bold** indicates the best values. '-'→Not applicable.

Model	AIWD16 (Kondapally et al., 2026)							
	BLEU-4		ROUGE-1		ROUGE-2		ROUGE-L	
	Caption	VQA	Caption	VQA	Caption	VQA	Caption	VQA
BLIP (Li et al., 2022)	39.52	34.44	38.63	36.05	27.34	24.34	37.55	30.55
BLIP-2 (Li et al., 2023a)	43.69	41.22	44.95	42.48	31.93	29.18	43.10	38.40
$EM-VLM4AD_{base}$ (Gopalkrishnan et al., 2024)	38.54	39.27	39.23	45.61	26.66	28.79	38.88	39.05
$EM-VLM4AD_{Q-large}$ (Gopalkrishnan et al., 2024)	43.59	37.95	45.61	44.23	28.79	26.66	43.12	38.88
DriveLM-Agent (Sima et al., 2025)	45.12	39.99	47.36	43.67	29.55	27.45	42.31	38.69
TWRFormer (Ours)	**47.13**	**41.83**	**49.29**	**45.80**	**31.99**	**29.97**	**43.77**	**39.98**

Table 12: Performance comparison between the proposed and state-of-the-art models for the image captioning and VQA tasks on the DriveLM and Nuscenes datasets. **Bold** indicates the best values. '-'→Not applicable.

Model	DriveLM				Nuscenes			
	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
BLIP (Li et al., 2022)	37.16	38.01	26.23	35.97	35.25	35.23	27.06	39.63
BLIP-2 (Li et al., 2023a)	41.23	42.69	29.87	38.98	41.11	41.88	27.79	36.54
$EM-VLM4AD_{base}$ (Gopalkrishnan et al., 2024)	45.36	57.77	29.11	51.98	42.36	-	-	66.96
$EM-VLM4AD_{Q-large}$ (Gopalkrishnan et al., 2024)	40.11	55.91	27.85	55.72	-	-	-	-
DriveLM-Agent (Sima et al., 2025)	53.09	56.11	28.99	56.79	-	-	-	-
LLAMA-Adapter (Zhang et al., 2024)	-	-	-	-	43.17	-	-	65.09
Omni (Wang et al., 2025)	-	-	-	-	42.45	-	-	61.55
TWRFormer (Ours)	**53.81**	**58.04**	**29.15**	**62.45**	**45.73**	**59.95**	**28.56**	**67.56**

7 Ablation study

Table 13 presents an ablation analysis on the AIWD16 dataset, evaluating the impact of key components in TWRFormer. Each variant modifies one element, baseline perception without restoration, transformer backbone with a convolutional tail (no encoder/decoder depth (M,N)), standard convolution in place of DSC, removal of weather queries, and full model without perceptual loss while keeping all other settings fixed. We report *PSNR*, *SSIM*, *mAP*, *mIoU*, and *BLEU* (VQA) to capture both restoration quality and downstream perception performance.

This modular evaluation highlights how each design choice influences both low-level restoration and high-level perception. To ensure controlled comparisons, we systematically disable or modify individual components while keeping all other settings fixed. ***(i)*** **No restoration (baseline):** evaluating perception directly on degraded inputs establishes a lower bound, yielding poor performance (31.2 *mAP*, 41.35 *mIoU*, 32.05 *BLEU*)

Table 13: Ablation study of the proposed TransWeatherNet on AIWD16 (Kondapally et al., 2026) to find the performance of each component. w/o→Without, Wq→weather queries, PL→perceptual loss.

Model	PSNR	SSIM	mAP	mIoU	BLEU
No restoration	-	-	31.2	41.35	32.05
Transformer base	24.28	0.832	41.6	49.91	38.32
TWRFormer w/o Wq	25.48	0.841	44.1	53.77	40.78
TWRFormer w/o DSC	26.81	0.853	47.5	57.52	43.93
Full TWRFormer w/o PL	27.97	0.875	48.1	58.25	45.22
Full TWRFormer + PL	**29.82**	**0.892**	**49.5**	**60.40**	**47.13**

and confirming the severe impact of transitional weather. ***(ii)* Transformer base:** replacing the proposed architecture with a shallow transformer captures limited global context but lacks hierarchical feature abstraction and multi-level reasoning. ***(iii)* Without weather queries (w/o Wq):** removing weather-aware embeddings prevents adaptation to varying weather intensities, leading to weaker reconstruction (25.48 *PSNR*) and reduced detection (41.1 *mAP*) and segmentation (53.77 *mIoU*). ***(iv)* Without DSC:** substituting depthwise-separable convolutions with standard convolutions increases computational cost without improving performance, slightly degrading both restoration and downstream metrics. ***(v)* Without perceptual loss (w/o PL):** removing perceptual supervision causes the model to overemphasize pixel fidelity while failing to preserve structural and semantic consistency. ***(vi)* Full model:** the complete TWRFormer, integrating hierarchical design, weather-aware queries, DSC fusion, and perceptual loss, achieves the best performance, demonstrating that each component contributes to robust restoration and improved perception under transitional weather.

8 Limitations and Future Work

Despite the effectiveness of the TransWeatheNet, current work has certain limitations that we plan to address in future research.

1. **Restoration consistency and efficiency.** While TWRFormer achieves strong restoration across diverse and unseen weather datasets, it operates on single-frame inputs without explicitly enforcing temporal consistency, which may affect stability in sequential driving scenarios.
2. **Model robustness to adversarial perturbations.** Although the dataset includes a wide range of transitional weather conditions, heavy noise or adversarial disturbances can still degrade model performance and reliability when deployed in the real world. To strengthen the proposed model against these vulnerabilities, future work will investigate both adversarial attack methods and corresponding defenses, with the goal of making the model more robust to naturally occurring noise as well as intentional perturbations.
3. **Risk analysis.** Beyond robust restoration and scene perception, the next research direction involves embedding risk assessment within the perception pipeline. By identifying hazardous regions in real time, subsequent models will advance from semantic reasoning to context-aware risk localization, essential for ensuring safety in autonomous driving amid transitional weather. Pairing these capabilities with causal inference will deepen explainability and foster trust in automated systems.

9 Conclusion

In this work, we present TransWeatherNet, a modular two-stage framework that jointly models and mitigates the challenges of transitional weather restoration and perception for autonomous driving. The framework integrates TWRFormer, a transformer-based restoration network that models global spatiotemporal dependencies and localized degradations through adaptive attention mechanisms. Furthermore, we perform downstream perception tasks, including object detection, semantic segmentation, and vision-language modeling, enabling unified training, evaluation, and cross-modal consistency. Extensive experiments on AIWD16 and four other public benchmarks, Foggy Cityscapes, RainCityscapes, ACDC, and BDD100K, demonstrate that TransWeatherNet achieves consistent gains of +2.01% *PSNR*, +0.03% *SSIM*, +3.1% *mAP*, +3.5% *IoU*, and +2.01% *BLEU*, establishing new state-of-the-art performance in both restoration and perception. By bridging restoration, and scene understanding, TransWeatherNet provides a unified framework for robust autonomous vision in dynamic weather conditions. Future work will focus on improving temporal consistency, physical realism, and computational efficiency for real-time deployment, as well as exploring multimodal extensions (e.g., LiDAR-weather fusion) on ADP platforms.

References

David Brüggemann, Christos Sakaridis, Tim Brödermann, and Luc Van Gool. Contrastive model adaptation for cross-condition robustness in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11378–11387, 2023.

Holger Caesar, Vikash Bankiti, Simon Hodge, Shenghong Liao, Xin Lyu, Vibhav Singh, Shaoshuai Xu, Zhiqiang Zhang, and Lijun Zhang. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1160–1170, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Haofeng Chen, Qi Dou Zhang, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Transunet: Transformers make strong encoders for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 407–418. Springer, 2021.

I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, et al. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17969–17979, 2025a.

I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17969–17979, 2025b.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018a.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022a.

Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17653–17662, 2022b.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339, 2018b.

Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15413–15423, 2023.

B. Dwyer, J. Nelson, T. Hansen, and et al. Roboflow (version 1.0). Software, 2025. URL `https://roboflow.com`. Computer vision.

Zaid A El-Shair, Abdalmalek Abu-raddaha, Aaron Cofield, Hisham Alawneh, Mohamed Aladem, Yazan Hamzeh, and Samir A Rawashdeh. Sid: Stereo image dataset for autonomous driving in adverse conditions. In *NAECON 2024-IEEE National Aerospace and Electronics Conference*, pp. 403–408. IEEE, 2024.

Vijay M. Galshetwar, Praful Hambarde, Prashant W. Patil, Akshay Dudhane, and Sachin Chaudhary. Clear roads, clear vision: Advancements in multi-weather restoration for smart transportation. *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–20, 2026.

Akshay Gopalkrishnan, Ross Greer, and Mohan Trivedi. Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. *arXiv:2403.19838*, 2024.

Yubin Gu, Yuan Meng, Jiayi Ji, and Xiaoshuai Sun. Acl: Activating capability of linear attention for image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17913–17923, 2025.

Xuyi He, Yuhui Quan, Ruotao Xu, and Hui Ji. A universal scale-adaptive deformable transformer for image restoration across diverse artifacts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12731–12741, 2025.

Mazin Hnewa and Hayder Radha. Integrated multiscale domain adaptive yolo. *IEEE Transactions on Image Processing*, 32:1857–1867, 2023.

Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 749–757, 2020.

Haofeng Hu, Liang Chen, Zhongyuan Wang, Lei Zhu, Mengyang Zhang, and Yun Fu. Depth-attentional features for single-image rain removal. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8022–8031, 2019.

JiaKui Hu, Zhengjian Yao, Lujia Jin, Hangzhou He, and Yanye Lu. Enhancing image restoration transformer via adaptive translation equivariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16047–16057, October 2025.

Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing*, 30:1759–1770, 2021.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-end multimodal model for autonomous driving. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=kH3t5lmOU8`.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Vaishnavi Khindkar, Chetan Arora, Vineeth N Balasubramanian, Anbumani Subramanian, Rohit Saluja, and CV Jawahar. To miss-attend is to misalign! residual self-attentive feature alignment for adapting object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3632–3642, 2022.

Madhavi Kondapally, K Naveen Kumar, and C Krishna Mohan. Object detection in transitional weather conditions for autonomous vehicles. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024a. doi: 10.1109/IJCNN60899.2024.10651445.

Madhavi Kondapally, K. Naveen Kumar, Chalavadi Vishnu, and C. Krishna Mohan. Towards a transitional weather scene recognition approach for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):5201–5210, 2024b. doi: 10.1109/TITS.2023.3331882.

Madhavi Kondapally, K Naveen Kumar, and C Krishna Mohan. Eyes on the road, words in the changing skies: Vision-language assistance for autonomous driving in transitional weather. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=PCEDvdVJon`.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 12809–12819, June 2025a.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12809–12819, 2025b.

Fanjie Kong, Yitong Li, Weihuang Chen, Chen Min, Yizhe Li, Zhiqiang Gao, Haoyang Li, Zhongyu Guo, and Hongbin Sun. Vlr-driver: Large vision-language-reasoning models for embodied autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26966–26976, October 2025c.

Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19994–20006, 2023.

Ashutosh Kulkarni, Prashant W. Patil, Santosh Kumar Vipparthi, Subrahmanyam Murala, and Balasubramanian Raman. Dtmir-pro: Domain translation with prompt-based latent-space generalization for multi-weather image restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3846–3856, March 2026.

Jianyu Lai, Sixiang Chen, Yunlong Lin, Tian Ye, Yun Liu, Song Fei, Zhaohu Xing, Hongtao Wu, Weiming Wang, and Lei Zhu. Snowmaster: Comprehensive real-world image desnowing via mllm with multi-model feedback optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4302–4312, June 2025.

Sohyun Lee, Taeyoung Son, and Suha Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18911–18921, 2022.

Sam Leroux, Bo Li, and Pieter Simoens. Multi-branch neural networks for video anomaly detection in adverse lighting and weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2358–2366, 2022.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, Shuchang Zhou, Li Zhang, Xiaojuan Qi, Hao Zhao, Mu Yang, Wenjun Zeng, and Xin Jin. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 11971–11981, June 2025a.

Jianping Li, Qifan Tan, Songchao Tan, Xiao Ke, Zhiwei Li, Tianyu Shen, and Guozhen Tan. Uniperception: Toward unification of perception using multi-stage training pipeline in adverse weather conditions. *IEEE Transactions on Intelligent Vehicles*, 11(3):415–427, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/li22n.html`.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Mingjia Li, Binhui Xie, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Vblc: Visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8605–8613, 2023b.

Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1633–1642, 2019.

Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3175–3185, 2020.

Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 254–269, 2018.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pp. 469–485. Springer, 2025b.

Xuelong Li, Kai Kou, and Bin Zhao. Weather gan: Multi-domain weather translation using generative adversarial networks. *arXiv preprint arXiv:2103.05422*, 2021.

Jingyun Liang, Jiezhang 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.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 12037–12047, June 2025.

Hongbin Lin, Yiming Yang, Chaoda Zheng, Yifan Zhang, Shuaicheng Niu, Zilu Guo, Yafeng Li, Gui Gui, Shuguang Cui, and Zhen Li. Driveflow: Rectified flow adaptation for robust 3d object detection in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6943–6951, 2026.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pp. 430–448. Springer, 2024.

Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22369–22380, 2025.

Taorong Liu, Jing Xiao, Liang Liao, and Chia-Wen Lin. Towards robust online domain adaptive semantic segmentation under adverse weather conditions. *arXiv preprint arXiv:2409.01072*, 2024.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pp. 21–37. Springer, 2016.

Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1792–1800, 2022.

Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7007–7016, 2019.

Yidi Liu, Dong Li, Xueyang Fu, Xin Lu, Jie Huang, and Zheng-Jun Zha. Uhd-processer: Unified uhd image restoration with progressive frequency learning and degradation-aware prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23121–23130, June 2025.

Yuetong Liu, Yunqiu Xu, Yang Wei, Xiuli Bi, and Bin Xiao. Clear nights ahead: Towards multi-weather nighttime image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7395–7403, 2026.

Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018.

Xin Lu, Jie Xiao, Yurui Zhu, and Xueyang Fu. Continuous adverse weather removal via degradation-aware distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 28113–28123, June 2025a.

Xin Lu, Jie Xiao, Yurui Zhu, and Xueyang Fu. Continuous adverse weather removal via degradation-aware distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28113–28123, 2025b.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12764–12777, June 2025a.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12764–12777, 2025b.

Kondapally Madhavi and C Krishna Mohan. Weather scene perception for autonomous vehicles. *IEICE Proceeding Series*, 81:61–64, 2024.

Kondapally Madhavi, K Naveen Kumar, and C Krishna Mohan. Transwardx: An explainable black-box object detection attack for autonomous driving in transitional weather conditions. In *International Conference on Pattern Recognition*, pp. 275–290. Springer, 2024.

Kondapally Madhavi, K Naveen Kumar, C Krishna Mohan, and Sobhan Babu. Cars: A causal intervention segmentation framework and benchmark dataset for autonomous driving under transitional weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4046–4056, March 2026a.

Kondapally Madhavi, K. Naveen Kumar, and C. Gayathri. Twfnet: Introducing transitional weather conditions for autonomous driving with a spatio-temporal forecasting network. *Pattern Recognition*, 171:112154, 2026b. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112154. URL `https://www.sciencedirect.com/science/article/pii/S0031320325008143`.

Debabrata Mandal, Soumitri Chattopadhyay, Guansen Tong, and Praneeth Chakravarthula. Unicorn: Latent diffusion-based unified controllable image restoration network across multiple degradations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1543–1553, March 2026.

Aboli Marathe, Deva Ramanan, Rahee Walambe, and Ketan Kotecha. Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3318–3327, 2023.

Valentina Mușat, 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.

Kemal Oksuz, Alexandru Buburuzan, Anthony Knittel, Yuhan Yao, and Puneet K. Dokania. Foundation models for trajectory planning in autonomous driving: A review of progress and open challenges. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=E2L5J2O2Bk`. Survey Certification.

Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):10346–10357, 2023.

Yuwen Pan, Rui Sun, Wangkai Li, and Tianzhu Zhang. Exploring weather-aware aggregation and adaptation for semantic segmentation under adverse conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13952–13962, October 2025.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023.

Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2482–2491, 2018.

Haina Qin, Wenyang Luo, Libin Wang, Dandan Zheng, Jingdong Chen, Ming Yang, Bing Li, and Weiming Hu. Reversing flow for image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7545–7558, 2025.

Sudarshan Rajagopalan and Vishal M Patel. Awracle: All-weather image restoration using visual in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6675–6683, 2025.

Sudarshan Rajagopalan, Nithin Gopalakrishnan Nair, Jay N. Paranjape, and Vishal M. Patel. Gendeg: Diffusion-based degradation synthesis for generalizable all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28144–28154, 2025.

Qiyu Rong, Hongyuan Jing, Mengmeng Zhang, Jinlong Li, and Mengfei Han. Strrnet: Semantics-guided two-stage raindrop removal network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1370–1378, June 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 2015.

Thomas Rothmeier and Werner Huber. Let it snow: On the synthesis of adverse weather image data. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 3300–3306, 2021. doi: 10.1109/ITSC48978.2021.9565008.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10765–10775, 2021.

Tasneem Shaffee and Sherief Reda. Robumtl: Enhancing multi-task learning robustness against weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5591–5600, March 2026.

Yinzhe Shen, Omer Sahin Tas, Kaiwen Wang, Royden Wagner, and Christoph Stiller. Divide and merge: Motion and semantic learning in end-to-end autonomous driving. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=RvtCNm1Rdv`.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pp. 256–274. Springer, 2025.

Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *European Conference on Computer Vision*, pp. 749–765. Springer, 2020.

Huawei Sun, Zixu Wang, Hao Feng, Julius Ott, Lorenzo Servadei, and Robert Wille. TRIDE: A text-assisted radar-image weather-aware fusion network for depth estimation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=ZMqMnwMfse`.

Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *European Conference on Computer Vision*, pp. 111–129. Springer, 2024.

Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28165–28175, June 2025a.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025b.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5769–5780, 2022.

Tzutalin. Labelimg. Free Software: MIT License, 2015. URL `https://github.com/tzutalin/labelImg`.

Jeya Maria Jose Valanarasu, 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.

Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4526, 2021.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

Hong Wang, Zongsheng Yue, Qi Xie, Qian Zhao, Yefeng Zheng, and Deyu Meng. From rain generation to rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14791–14801, 2021a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22442–22452, 2025.

Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24090–24099, 2023.

Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9603–9612, 2021b.

Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12978–12995, 2022.

Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, Yang Zhou, Huaxiu Yao, and Zhengzhong Tu. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=z2VZl6sH7T`.

Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11724–11733, 2020.

Zhenhua Xu, Yan Bai, Yujia Zhang, Zhuoling Li, Fei Xia, Kwan-Yee K. Wong, Jianqiang Wang, and Hengshuang Zhao. Drivegpt4-v2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17261–17270, June 2025.

Fei Yang, Jianfeng Ren, Zheng Lu, Jialu Zhang, and Qian Zhang. Rain-component-aware capsule-gan for single image de-raining. *Pattern Recognition*, 123:108377, 2022.

Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24902–24912, June 2024.

Huitong Yang, Zhuoxiao Chen, Fengyi Zhang, Zi Huang, and Yadan Luo. Codemerge: codebook-guided model merging for robust test-time adaptation in autonomous driving. *Advances in Neural Information Processing Systems*, 38:49096–49123, 2026.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *The International Journal of Robotics Research*, 39(1):1–18, 2020.

Kejin Yu, Yuhan Sun, Taiqiang Wu, Ruixu Zhang, Zhiqiang Lin, Yuxin Meng, Junjie Wang, and Yujiu Yang. A survey of reasoning in autonomous driving systems: Open challenges and emerging paradigms. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=XwQ7dc4bqn`. Survey Certification.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

Syed Waqas Zamir, 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.

Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, 30:7419–7431, 2021.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024.

Xin Zhang and Robby T. Tan. Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 14527–14537, June 2025.

Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, and Dacheng Tao. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14781–14791, June 2024.

Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25445–25455, 2024.

Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, Jie Bai, Zhixiong Ma, and Xichan Zhu. Omnihd-scenes: A next-generation multimodal dataset for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):7032–7049, 2026.

Shen Zheng, Changjie Lu, Yuxiong Wu, and Gaurav Gupta. Sapnet: Segmentation-aware progressive network for perceptual contrastive deraining. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 52–62, 2022.

Mo Zhou, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Vishal M. Patel, and Hossein Talebi. Unires: Universal image restoration for complex degradations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13237–13247, October 2025.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13782–13790, 2026.

Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21747–21758, 2023.