RainDiffusion: A Diffusion Framework for Deraining

Ziyue Lin, Runshi Yang, and Yi Zhang

Abstract—Outdoor vision systems, ranging from everyday photography to safety-critical perception, need to operate reliably in rainy conditions, but even light rainfall can still drastically degrade image quality and confuse downstream algorithms. Rain streaks, raindrops, and mixed rain patterns introduce structured, spatially varying artifacts that reduce contrast, distort edges, and suppress fine textures, making them fundamentally different from simple Gaussian noise. In this work, we introduce RainDiffusion, a diffusion-based deraining framework that formulates rain removal as a conditional generative denoising problem. Instead of using hand-designed priors, RainDiffusion learns the distribution of clean images and progressively reconstructs rain-free content through a multi-step reverse diffusion process. To better localize and suppress rain artifacts, we incorporate a learnable spatial guidance module that produces soft rain-attention maps, which are injected into the denoising network to emphasize rain-corrupted regions while preserving clean background structures. Extensive experiments on synthetic and real-world rainy datasets show that RainDiffusion achieves competitive or superior performance compared with state-of-the-art deraining methods in terms of PSNR and SSIM. Visual results further demonstrate that the proposed approach effectively removes diverse rain patterns while maintaining sharp edges and realistic textures, highlighting diffusion models as a powerful and flexible paradigm for image deraining.

Index Terms—Computational Photography

1 Introduction

Removing rain from an image is a challenging problem: streaks, droplets, and mixed rain patterns degrade contrast, distort edges, and hide fine textures in a structured and spatially uneven way. Over the past decade, deraining methods have evolved along three main lines, namely convolutional networks, adversarial frameworks, and diffusion based models. Each line offers clear benefits but also reveals limitations that motivate our design.

Early CNN based approaches, such as residual U-Net variants [1], denoising autoencoders [2], and progressive refinement networks like PReNet [3], show that multi scale feature extraction with residual learning can remove moderate rain, but their limited receptive field and lack of explicit spatial modeling often leave residual artifacts or over smoothed details under complex or mixed rain. To improve perceptual quality, GAN based methods such as U-Net generators with adversarial loss [4] and AttentiveGAN [5] introduce adversarial objectives and attention modules to highlight rain affected regions and sharpen textures, but unstable training and the focus on visual realism can cause hallucinated structures and geometric distortions, which is problematic when accurate reconstruction is required. Diffusion based generative models including DDPM and DDIM [6], [7] provide a stable optimization framework and strong modeling capacity by viewing restoration as progressive noise prediction, yet their use in deraining remains limited because the Gaussian noise assumption and nearly uniform denoising conflict with the structured, directional, and type dependent nature of real rain, and existing diffusion based restoration methods rely on global pixel wise or perceptual losses without explicit mechanisms to capture spatial heterogeneity or adapt to different rain types.

To address these limitations, we propose an attentive diffusion model that couples spatial attention with conditional diffusion for targeted rain removal. A dedicated attention mechanism predicts soft rain masks that are injected into the diffusion process to guide the denoising trajectory toward rain corrupted regions while preserving clean areas. This spatial conditioning relaxes the uniform denoising assumption of standard diffusion models and aligns the generative process with the geometry and structure of rain. In addition, we introduce variant embeddings that encode different rain types, including rain streaks, raindrops, and mixed patterns, which allows the model to adapt its denoising behavior to the observed rain characteristics. The proposed approach combines spatial attention, conditional diffusion, and type aware modulation to achieve precise and robust deraining, while retaining the computational cost of iterative diffusion sampling and relying on accurate attention masks for best performance.

Code: The code for this project is publicly available at GitHub.

2 Related Work

2.1 CNN-Based Image Restoration

Early deep learning approaches for single image deraining primarily relied on Convolutional Neural Networks (CNNs) to learn a direct mapping from rainy inputs to clean background scenes. Fu et al. introduced the Deep Detail Network (DDN) [8] which decomposed images into base and detailed layers, then use a ResNet architecture to remove rain streaks from the high-frequency detailed layer. Following this, recursive and multi-stage architectures became prominent. For instance, PRNet employed a multi-stage recursive network that repeatedly unfolded a shallow ResNet to separate rain streaks from the background. These CNN-based models offer strong representational capacity. However, they often struggle with heavy rain occlusion and tend to produce over-smoothed results due to the reliance on pixel-wise reconstruction losses.

• *Z. Lin, R.Yang and Y.Zhang are with the Department of Computer Science, University of Toronto.*

2.2 GAN-Based Deraining and Attention Mechanisms

To overcome the over-smoothing problems of CNNs, Generative Adversarial Networks (GANs) were introduced to improve the perceptual quality of restored images. Pix2Pix [9] pioneered a general framework image-to-image translation using GAN. By employing a conditional GAN (cGAN) architecture trained with both an adversarial loss and a pixel-wise L1 loss, Pix2Pix demonstrated that photorealistic restoration could be achieved. However, generic translation frameworks treat all pixels equally which fail to account for the spatial non-uniformity of rain.

To address this, Attentive GAN integrated visual attention mechanisms into the adversarial process. Their method employs an attentive module that explicitly predicts a attention map for raindrop region. Guided by this attention map, the generator operates in two stages: the first reconstructs structural content in degraded areas, while the second refines fine textures and restores global visual coherence. This spatial attention map guided generative restoration process is a primary inspiration for our proposed diffusion framework.

2.3 Diffusion Models for Image Restoration

Denoising Diffusion Probabilistic Models (DDPMs) have recently emerged as a powerful class of generative models, achieving state-of-the-art performance in various low-level vision tasks. Unlike CNNs or GANs, DDPMs formulate image restoration as a multi-step stochastic process that iteratively reverses a gradual noising chain to refine a random noise sample into a clean image. Diffusion models have shown remarkable success in general image restoration tasks like super-resolution [10] and inpainting [11]. However, their dedicated application to deraining remains relatively underexplored compared to other restoration domains, largely because classical diffusion models rely on the assumption that degradations follow additive Gaussian noise, which mismatches the highly structured and spatially heterogeneous nature of rain streaks and raindrops and thus limits their direct applicability without additional conditioning mechanisms.

3 Explore Data Analysis

In our exploratory data analysis (Figure 1 2 3 4), we investigated the behavior of high-pass and low-pass filtered images, along with different color spaces (RGB, HSV, and YCbCr), to better understand how raindrop and rain-streak noise affects image frequency components. The results show that high-pass regions, which encode edges and fine textures, are relatively robust to these degradations, whereas low-pass components, which capture smooth regions and global structure, suffer much more noticeable distortion. In the RGB and HSV spaces, the impact of noise remains fairly consistent across all channels, suggesting that color information is not the primary driver of noise sensitivity. In contrast, within the YCbCr space, the luminance (Y) channel is substantially more affected than the chrominance channels (Cb and Cr), indicating that raindrop and streak artifacts primarily perturb brightness and structural information rather than color variations.

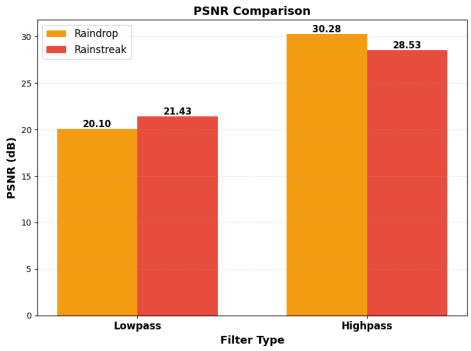

Fig. 1: High-pass and low-pass analysis

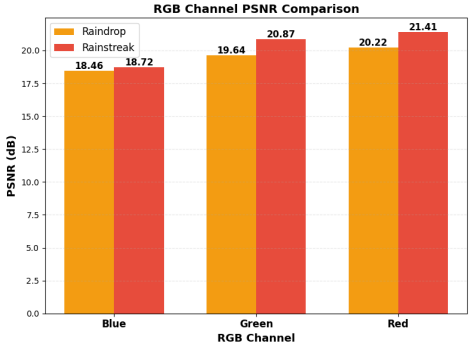

Fig. 2: RGB channels analysis

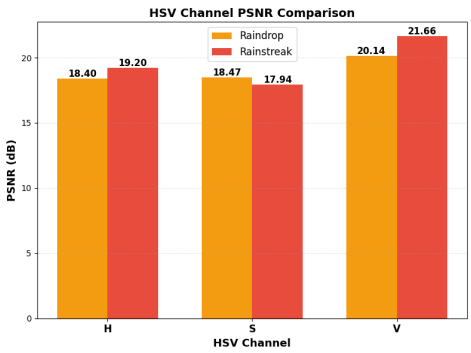

Fig. 3: HSV channels analysis

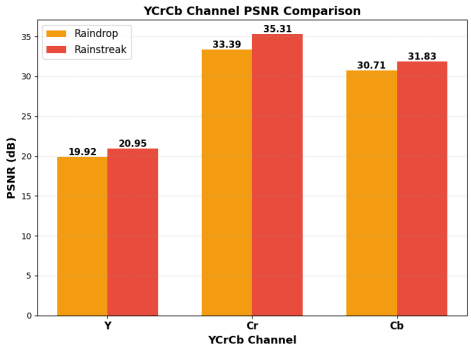

Fig. 4: YCrCb channels analysis

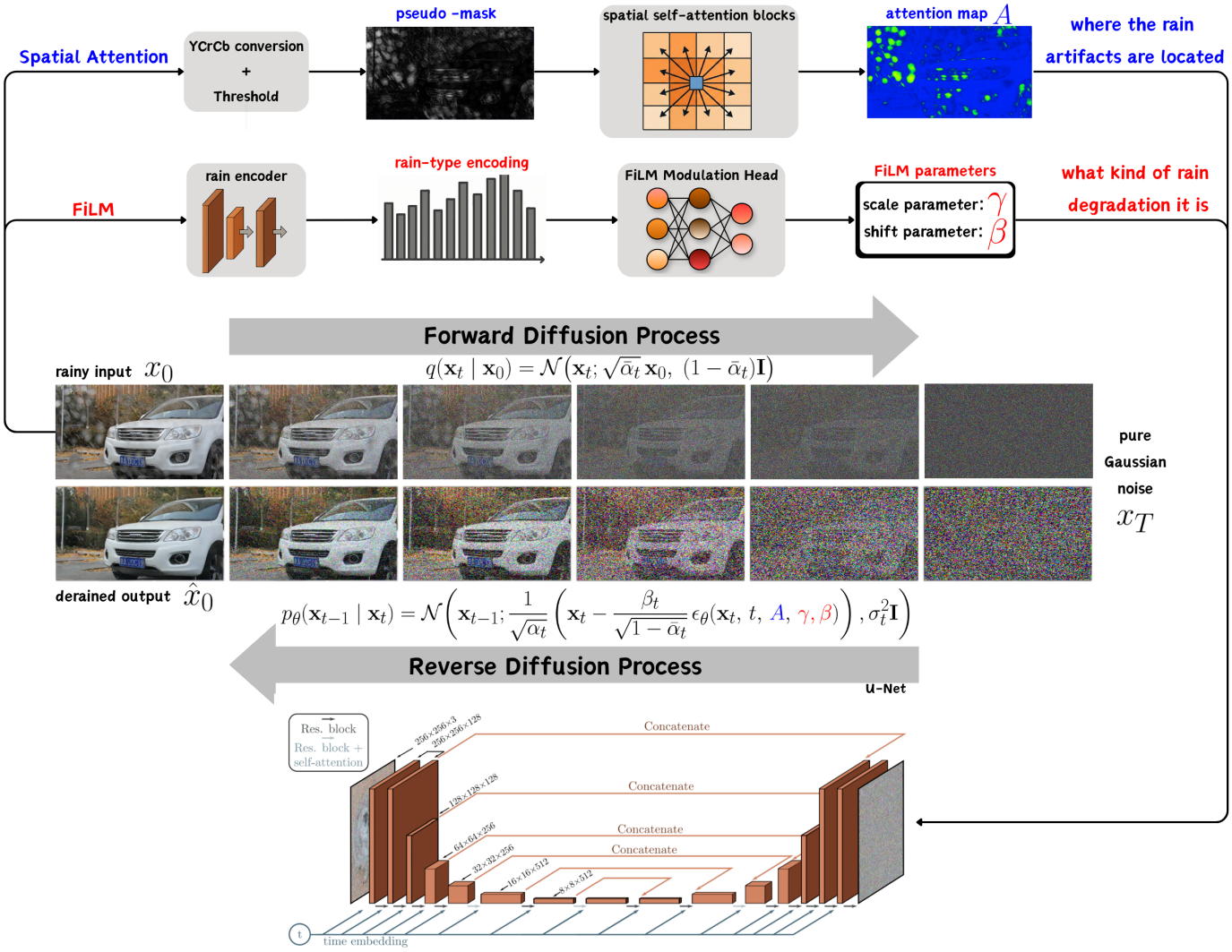

Fig. 5: Model Design

4 METHOD

We propose **RainDiffusion**, a conditional diffusion framework that integrates (1) a Spatial Attention module to locate rain-affected regions and (2) a FiLM-based modulation pathway to encode rain-type semantics into the reverse diffusion process. This design resolves the fundamental mismatch between classical diffusion (which assumes Gaussian noise) and the highly structured, heterogeneous nature of real rain artifacts.

RainDiffusion enhances the U-Net denoiser with both spatially varying and type-aware conditioning signals, enabling the reverse diffusion trajectory to target localized degradations while adapting to different rain categories. Below, we introduce each component and describe how they integrate into the forward and reverse diffusion processes.

4.1 Spatial Attention

EDA shows that luminance distortions (Y channel) strongly correlate with rain artifacts. We leverage this by generating a soft pseudo-mask from luminance difference ΔY. High-ΔY regions receive larger weights, producing a continuous-valued prior rather than a binary segmentation.

A trainable Spatial Attention block refines this luminance-derived prior. Given the rainy input image, we construct the query Q, key K, and value V tensors by applying three separate learned linear projections to the **same** rainy image. The attention map is computed using a standard self-attention formulation:

$$\mathbf{A} := \mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

Here, d denotes the embedding dimensionality used to scale the dot-product attention. And the resulting attention map $\mathbf{A}$ highlights rain-damaged pixels.

During reverse diffusion, $\mathbf{A}$ is injected as an additional conditioning feature at every scale of the U-Net, guiding denoising capacity toward corrupted regions. To remain consistent with the luminance prior, A is supervised with BCE + Dice losses.

4.2 Feature-wise Linear Modulation (FiLM)

To model rain-type semantics (raindrop, streak, mixture), we encode the rain category and combine it with timestep

embeddings through an MLP. This produces FiLM parameters: scale γ, and shift β, which modulate intermediate feature maps h via:

$$\text{FiLM}(h) = \gamma \odot h + \beta.$$

FiLM allows the denoiser to adjust its behavior based on both timestep and rain-type information, making the reverse process *type-aware* rather than uniform across degradations.

4.3 RainDiffusion U-Net Architecture

The backbone of RainDiffusion is a U-Net with residual blocks, self-attention, and skip connections. At each resolution, the feature update incorporates:

1) Spatial Attention map $\mathbf{A}$ (location of degradations),
2) FiLM modulation parameters (γ, β) (type of degradation),
3) Timestep embeddings,
4) Skip connections for multi-scale structure recovery.

Thus, the U-Net predicts the forward-noise residual $\epsilon_\theta(x_t, t, A, \gamma, \beta)$ conditioned on both *where* degradation occurs and *what kind* of rain is present as the mechanisms illustrated in Figure 5.

4.4 Forward Diffusion Process

The forward diffusion process gradually corrupts the clean image x_0:

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t;\ \sqrt{\bar{\alpha}_t}\, x_0,\ (1 - \bar{\alpha}_t)\, I\big)\,,$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$ denotes the accumulated product of the noise attenuation coefficients. This formulation follows the standard forward process used in DDPMs.

4.5 Reverse Diffusion Process (RainDiffusion)

In contrast, the reverse process is *conditional*. At timestep t, the conditional reverse transition is:

$$p_\theta(x_{t-1} \mid x_t, A, \gamma, \beta) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t, A, \gamma, \beta),\ \sigma_t^2 I\big)\,,$$

where the mean is defined as:

$$\mu_\theta(x_t, t, A, \gamma, \beta) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\, \epsilon_\theta(x_t, t, A, \gamma, \beta) \right).$$

This explicitly incorporates:

- $\mathbf{A} \rightarrow$ where rain artifacts are located
- $\gamma, \beta \rightarrow$ what type of rain it is
- $t \rightarrow$ how much noise is present

thus making denoising both *spatially selective* and *semantically adaptive*.

4.6 Loss Function

Our training objective integrates four complementary loss terms to jointly guide noise estimation, image reconstruction, sample consistency, and mask supervision. The overall loss is formulated as:

$$\mathcal{L} = \lambda_{\text{eps}}\mathcal{L}_{\text{eps}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}.$$

Noise Loss $\mathcal{L}_{\text{eps}}$: This term enforces accurate Gaussian noise prediction by minimizing the ℓ_2 distance between the predicted noise $\hat{\epsilon}$ and the true diffusion noise ϵ:

$$\mathcal{L}_{\text{eps}} = \|\hat{\epsilon} - \epsilon\|_2^2.$$

It ensures the denoising model faithfully aligns with the diffusion forward process.

Reconstruction Loss $\mathcal{L}_{\text{rec}}$: The ℓ_1 reconstruction loss encourages the restored clean image $\hat{x}_0$ to match the ground-truth clean target x_0:

$$\mathcal{L}_{\text{rec}} = \|\hat{x}_0 - x_0\|_1.$$

This improves perceptual fidelity and enhances structural recovery.

Consistency Loss $\mathcal{L}_{\text{cons}}$: To reduce stochastic variance during diffusion sampling, the consistency loss penalizes discrepancies between two independently sampled predictions $\hat{x}_0^{(i)}$ and $\hat{x}_0^{(j)}$:

$$\mathcal{L}_{\text{cons}} = \|\hat{x}_0^{(i)} - \hat{x}_0^{(j)}\|_1.$$

This regularizes training and improves output stability.

Segmentation Loss $\mathcal{L}_{\text{seg}}$: The mask prediction module is supervised using a combination of BCE and Dice losses:

$$\mathcal{L}_{\text{seg}} = \text{BCE}(m_{\text{pred}}, m_{\text{pseudo}}) + \text{Dice}(m_{\text{pred}}, m_{\text{pseudo}}).$$

This pseudo-mask guided supervision improves rain-region localization and provides robustness to foreground-background imbalance.

5 Experimental Results

5.1 Datasets

We conduct experiments on the RainDS dataset (Figure 6) introduced by Quan et al. [12], which provides paired rainy–clean images covering three degradation types: raindrop, rainstreak, and mixed rain. The dataset contains approximately 700 aligned pairs, with about 220 samples per category. RainDS includes both synthetic and real captures across diverse scenes, offering substantial variation in lighting, textures, background clutter, and rain intensity. All images have consistent spatial alignment between the rainy and ground-truth clean versions, making the dataset suitable for supervised deraining. We use the full dataset for training and testing and report results separately for each rain type to highlight performance differences across degradation modes.

Ground Truth **Raindrop Degradation** **Rainstreak Degradation**

Fig. 6: Examples from the RainDS dataset

TABLE 1: Training hyperparameters and configurations

Parameter	Value
Epochs	300
Batch size	4
Learning rate	0.0002
Optimizer	AdamW
AdamW β_1	0.9
AdamW β_2	0.999
Reconstruction loss weight (λ_{rec})	0.1
Consistency loss weight (λ_{cons})	0.005
Image size	256×256
Diffusion timesteps (T)	1000
Noise schedule start (β_{start})	0.0001
Noise schedule end (β_{end})	0.02
EMA decay rate	0.999
Evaluation interval	Every 10 epochs

5.2 Training Specifications

We train all models using the AdamW optimizer with a fixed learning rate, and all images are resized to a resolution of 256×256. Our training objective combines a mean squared error noise-prediction loss to supervise diffusion denoising, an L1 reconstruction loss weighted by λ_{rec} to encourage accurate recovery of clean structures, and a lightly weighted pairwise consistency term λ_{cons} that promotes structural coherence across samples within the same batch. To improve computational efficiency, we train with automatic mixed precision and maintain an exponential moving average (EMA) of model parameters to stabilize learning. At each iteration, timesteps are uniformly sampled from a 1000-step linear noise schedule. We periodically perform full denoising inference on the validation set and compute PSNR and SSIM for monitoring training progress. Model checkpoints are saved based on the best reconstruction performance, and both the final and EMA-smoothed weights are retained for evaluation.

5.3 Comparison Methods and Evaluation Metrics

We compare RainDiffusion against three representative baselines commonly used in image deraining:

- **UNet-CNN**: a convolutional restoration model serving as a classical feed-forward baseline.
- **AttentiveGAN**: an adversarial approach enhanced with spatial attention for perceptual detail.
- **Vanilla Diffusion**: a DDPM-based denoising model without attention or FiLM conditioning.

To provide a comprehensive evaluation, we assess performance across three complementary dimensions:

1) **Qualitative rain removal** — visual comparisons on challenging raindrop, rainstreak, and mixed-rain scenes.
2) **Quantitative fidelity** — PSNR and SSIM measuring pixel-level accuracy and structural preservation.
3) **Downstream robustness** — lane detection performance on ground-truth, rainy, and derained images.

These three perspectives jointly evaluate how well each model restores clean structure, preserves perceptual detail, and supports real-world vision tasks under adverse weather.

5.3.1 Qualitative Results

Figure 7 presents qualitative comparisons across three representative rain types. UNet-CNN removes coarse rain patterns but tends to over-smooth fine textures, especially along thin structures such as lane markings or foliage. AttentiveGAN produces sharper textures but occasionally introduces hallucinated edges due to adversarial instability. The diffusion baseline offers cleaner global structure yet struggles with small, spatially localized artifacts such as droplets. In contrast, our attentive RainDiffusion model consistently handles both localized and linear rain degradations, preserving sharper boundaries and producing visually stable results.

5.3.2 Quantitative Results

RainDiffusion achieves the highest or second-highest scores across all subsets, outperforming both CNN and GAN baselines while providing more balanced performance than the vanilla diffusion model. Notably, our method excels in raindrop and mixed-rain scenarios, where localized distortions require targeted spatial attention; this is exactly what the proposed attention modules are designed to capture. The diffusion baseline performs strongly on rainstreak scenes due to the more global and directional nature of streak artifacts but lacks the localized guidance needed for other rain types, explaining its diminished performance in the remaining categories.

TABLE 2: PSNR (dB) comparison across three rain types.

Method	Raindrop	Rainstreak	Mixture
UNet-CNN	20.15	21.74	18.88
AttentiveGAN	21.60	23.77	20.72
Diffusion (no attention)	22.46	**24.66**	21.05
RainDiffusion	**22.57**	24.46	**21.61**

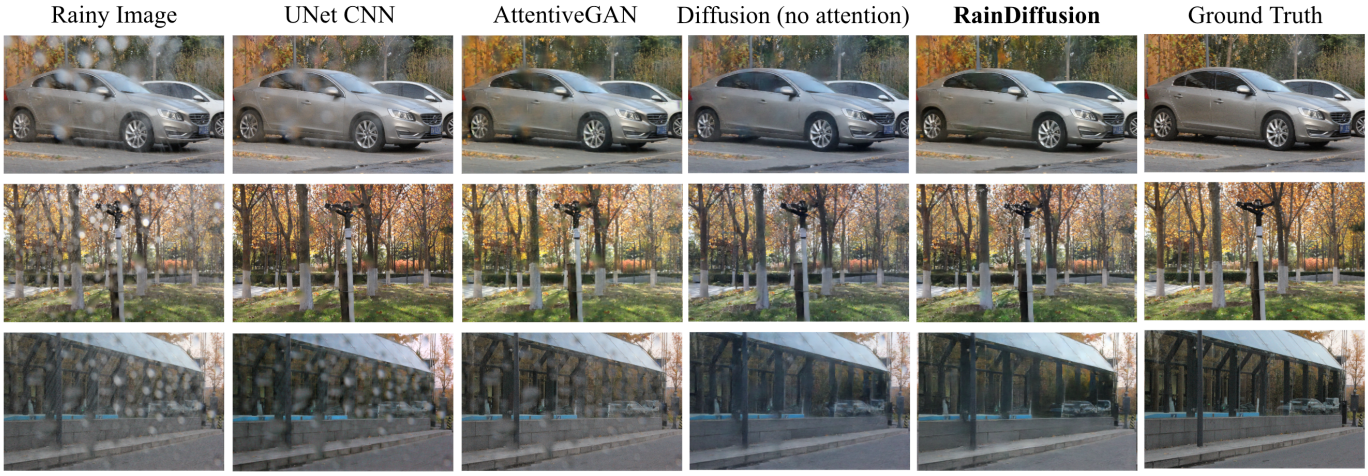

Fig. 7: Qualitative comparison across raindrop, rainstreak, and mixed rain conditions.

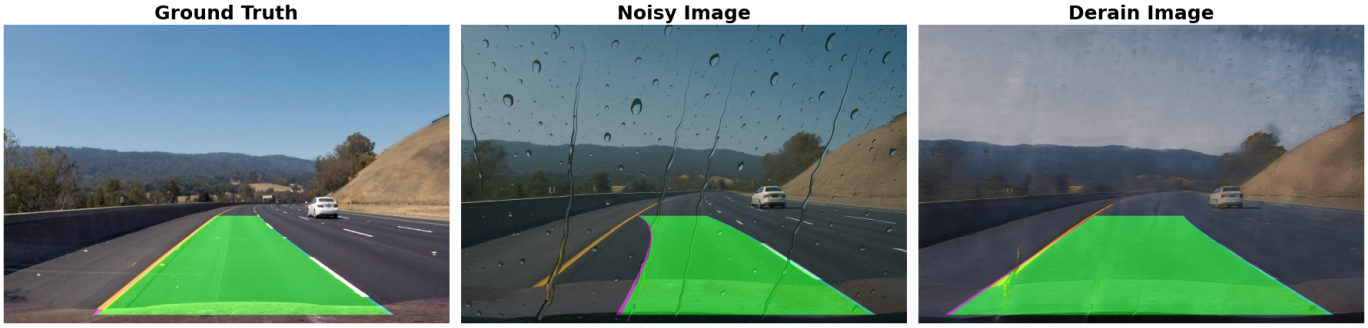

Fig. 8: Lane detection results on clean, rainy, and derained images.

TABLE 3: SSIM comparison across three rain types.

Method	Raindrop	Rainstreak	Mixture
UNet-CNN	0.5768	0.6108	0.5229
AttentiveGAN	0.7019	0.7561	0.6449
Diffusion (no attention)	0.7255	**0.7858**	0.6690
RainDiffusion	**0.7378**	0.7723	**0.6808**

5.3.3 *Downstream Evaluation: Lane Detection*

To evaluate real-world utility, we further assess model performance on a downstream perception task—lane detection. Rainy images significantly degrade detection quality due to occlusions and contrast suppression. As shown in Figure 8, detections on rainy inputs are fragmented and misaligned, especially at long range.

Deraining with attentive RainDiffusion recovers continuous lane boundaries and reduces false positives along road edges. Its outputs are visually and functionally closer to ground-truth detections than those produced by competing models, demonstrating improved robustness for downstream tasks operating under adverse weather.

5.4 Discussion

Our results reveal several design trade-offs. Spatial attention significantly improves the removal of localized artifacts such as droplets and helps avoid over-smoothing in clean regions, while the variant embeddings enhance robustness by allowing the model to adapt to different rain types. These benefits, however, come at the cost of increased computational cost, as the attention module adds complexity and the multi-step diffusion process is inherently slower than single-pass CNN.

We also observe a few failure cases. In scenes with extremely dense mixed rain or strong specular highlights, the model may under-estimate the rain intensity or introduce mild texture softening. These issues are largely attributable to the reliance on learned attention masks, which may require stronger supervision or additional training to fully capture highly complex degradations.

6 Conclusion

In summary, our study demonstrates that integrating attention mechanisms into diffusion-based restoration framework yields substantial improvements for image deraining. Attentive RainDiffusion achieves cleaner structural recovery, sharper textures, and fewer residual artifacts compared with CNN, GAN, and vanilla diffusion model. Quantitative results further validate these observations, with our method attaining the highest PSNR and SSIM across most rain categories. The combination of spatial attention with feature-wise modulation enables the model to selectively emphasize rain-affected regions while preserving clean background details.

Despite these advantages, several limitations remain. The iterative nature of diffusion introduces higher compu-

tational cost than single-pass CNN, and the effectiveness of the method depends on the quality of the learned attention masks, which may underperform in scenes with extremely dense or visually ambiguous rain. These observations highlight opportunities for future work, such as developing more efficient sampling strategies, integrating lightweight attention predictors. Extending the framework to video deraining or multimodal sensing systems like, LiDAR–camera fusion, also represents a promising direction, enabling more reliable perception.

Overall, our results suggest that attention-enhanced diffusion models provide a compelling path forward for robust image restoration under adverse weather conditions. Beyond deraining, the principles explored in this work, spatially guided diffusion, adaptive conditioning, and variant-aware modulation, may generalize to other real-world degradation scenarios, offering a foundation for broader advances in weather-robust visual perception.

REFERENCES

[1] J. Gurrola-Ramos, O. Dalmau, and T. E. Alarcon, "A residual dense u-net neural network for image denoising," *IEEE Access*, vol. 9, pp. 31 742–31 754, 2021.

[2] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th international conference on data mining workshops (ICDMW)*. IEEE, 2016, pp. 241–246.

[3] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3937–3946.

[4] H. Porav, T. Bruls, and P. Newman, "I can see clearly now: Image restoration via de-raining," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7087–7093.

[5] Z. Yuan, Y. Zhao, Y. Wang, and X. Zuo, "Design and implementation of single image rain removal system based on deep learning," in *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*. IEEE, 2024, pp. 1205–1208.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*.

[8] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing*, vol. 26, no. 6, p. 2944–2956, Jun. 2017. [Online]. Available: http://dx.doi.org/10.1109/TIP.2017.2691802

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[10] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," 2021. [Online]. Available: https://arxiv.org/abs/2104.07636

[11] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[12] R. Quan, X. Yu, Y. Liang, and Y. Yang, "Removing raindrops and rain streaks in one go," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9147–9156.