Rain Removal via VAE-Enhanced Transformer with Hierarchical Feature Integration

YaYa Huang, LiTong Liu, and KokSheik Wong
Monash University Malaysia, Malaysia
E-mail: {yhua0261@student., lliu0118@student., wong.koksheik@}monash.edu

***Abstract*—Rain removal remains a challenging task in computer vision due to the complex interference of rain streaks with background structures. While Transformer-based methods have shown promising performances in modeling long-range dependencies, they have drawbacks such as a lack of prior guidance, confusion between rain textures and scene semantics, as well as insufficient cross-scale consistency. To address these limitations, we propose a novel image de-raining framework that integrates a Variational Autoencoder (VAE) with a Transformer backbone via a Latent-Guided Attention (LGA) mechanism. The VAE learns a global latent representation that captures the statistical structure of rain patterns, which is then injected into the Transformer at multiple levels through cross-attention modules. This enables the network to focus more effectively on essential scene features while suppressing rain artifacts. Furthermore, a hierarchical feature integration strategy is adopted to fuse low-level textures and high-level semantics across scales. Experiments demonstrate that our method achieves superior PSNR and SSIM scores on standard benchmark datasets, while maintaining robust performance under complex rain conditions. Our framework offers a new perspective on leveraging latent priors for enhancing Transformer-based image restoration, paving the way for future applications in image de-weathering and denoising.**

***Index Terms*—Image Rain Removal; Transformer; Latent-Guided Attention.**

I. Introduction

Image rain removal is a critical task in computer vision, aiming to restore clear background scenes from images degraded by rain streaks or droplets. With the rapid advancement of deep learning, this field has evolved from traditional physical priors, such as total variation-based denoising [1], to data-driven approaches based on convolutional neural networks (CNNs) [2] and Transformers [3]. However, existing methods still struggle with complex rain patterns, dynamic raindrops, and heavy rainfall, particularly when balancing fine-detail preservation with computational efficiency [4].

Specifically, CNN-based models such as the Deep Detail Network [6] suffer from limited receptive fields, which restrict their ability to model the long-range dependencies present in elongated rain streaks. On the other hand, Transformer-based frameworks such as the Image De-Raining Transformer (IDT) [5] effectively capture global semantic context but incur high computational costs, typically $\mathcal{O}(H^2W^2)$ for an input image of size $H \times W$, making them unsuitable for real-time applications. Furthermore, most existing methods only operate in the image space and do not explicitly leverage latent-space priors to separate rain layers from background textures [7].

In this work, we define **dense rain streaks** as high-frequency, spatially extensive rain patterns that severely obscure scene details and exhibit non-uniform motion blur. These rain structures pose significant restoration difficulty and often result in over-smoothed or artifact-laden outputs. To overcome these limitations, we propose a novel framework named **VTHFI** (VAE-Enhanced Transformer with Hierarchical Feature Integration). Our method combines the latent representation power of Variational Autoencoders (VAEs) with the global contextual modeling of Transformers. Specifically, we adopt a VAE-based encoder-decoder architecture that disentangles rain artifacts from semantic backgrounds by leveraging KL-divergence-regularized multi-scale latent embeddings [8]. To further enhance cross-scale information exchange, we design a bidirectional hierarchical backbone that dynamically integrates coarse-to-fine features, improving the model's capability to separate layered rain structures [9]. In addition, we incorporate lightweight local-window attention and frequency-domain convolutions to reduce inference cost while maintaining robust performance under diverse and complex rainy conditions [10].

Our work makes the following contributions: (a) We propose VTHFI, a unified framework that integrates variational latent modeling with Transformer-based global context reasoning for single-image rain removal; (b) We design a bidirectional hierarchical feature integration strategy to facilitate dynamic multi-scale feature fusion, enhancing rain-background separation and detail preservation; (c) We introduce efficient local attention and frequency-aware convolutional modules to improve inference efficiency without sacrificing restoration quality, and; (d) We conduct extensive experiments on five benchmark datasets, where VTHFI consistently outperforms state-of-the-art methods in terms of PSNR, SSIM, and visual fidelity, demonstrating superior generalization and robustness.

II. Related Work

The evolution of single-image de-raining methods can be categorized into three paradigms, namely: prior-based approaches, CNN-based models, and transformer-based frameworks.

979-8-3315-7206-8/25/$31.00

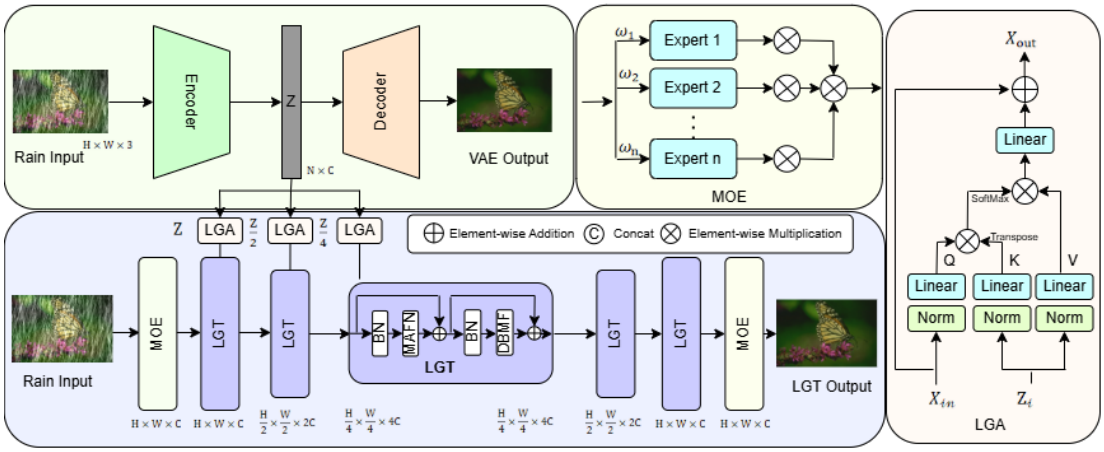


Fig. 1: Overall architecture of our proposed de-raining framework.

A. Prior-Based Methods

Early approaches relied on hand-crafted features derived from the physical properties of rain. For example, the Dark Channel Prior (DCP) [12] originally designed for haze removal was adapted for de-raining by exploiting the occlusion of dark channels caused by rain streaks. Luo et al. [2] proposed a Gaussian Mixture Model (GMM)-based method, leveraging discriminative sparse coding to separate rain from background. While these methods achieved reasonable performance in uniform rain scenarios, their reliance on fixed assumptions limited adaptability to complex real-world conditions.

B. CNN-Based Methods

The advent of deep learning revolutionized de-raining. For example, Fu et al. [6] introduced the Deep Detail Network (DDN), which directly mapped rainy images to clean ones via end-to-end learning. Subsequent works improved upon this by incorporating recursive structures (e.g., RESCAN [6]) and progressive refinement (e.g., PReNet [13]). Advanced architectures such as MSPFN [14] and RCDNet [15] employed multi-stream designs to capture diverse rain patterns. Recent CNN-based methods [16]–[18] further improved de-raining through multi-branch structures, graph-based reasoning, and spatial pyramid modules. However, CNNs inherently suffer from limited receptive fields, hindering their ability to model long-range dependencies, which are critical in dense rain scenarios.

C. Transformer-based Methods

Inspired by their breakthroughs in NLP, recent de-raining works have adopted transformers to capture the global context. For example, Uformer [19] introduced a U-shaped transformer for multi-scale feature extraction, while Restormer [9] optimized attention mechanisms for high-resolution restoration. Domain-specific designs such as DRSformer and LGAformer [20] further improved performance by integrating local-global attention. These methods outperform CNNs on benchmark datasets such as Rain200H and DID-Data [9], [20]. However, they often treat features uniformly, hence neglecting hierarchical relationships among scales, which limits fine-grained detail recovery.

In summary, while prior-based, CNN-based, and Transformer-based methods have progressively improved de-raining quality, significant limitations remain in global context modeling, hierarchical feature integration, and latent disentanglement. These gaps motivate our proposed approach. In the next section, we introduce a VAE-enhanced transformer framework that explicitly addresses these issues through generative guidance, multi-scale attention fusion, and adaptive specialization.

III. Methodology

We propose **VTHFI**, a VAE-Enhanced Transformer with Hierarchical Feature Integration, to tackle the problem of single image de-raining. Our design is motivated by three critical limitations observed in prior works: (1) inadequate global context modeling, (2) weak hierarchical feature fusion, and (3) the entanglement of rain streaks with background content. To address these challenges, the **VTHFI framework** integrates three key components: (1) a Variational Autoencoder (VAE) that extracts global semantic priors [8], (2) a Transformer-based restoration backbone termed **VAEformer** for multi-scale spatial representation [21], and (3) a Latent-Guided Attention (LGA) mechanism that injects semantic priors into the restoration pathway.

A. Overall Framework

Fig. 1 illustrates our proposed framework, which consists of three interacting components: a variational prior extraction branch, a hierarchical Transformer backbone (LGTformer), and a Mixture-of-Experts (MoE) routing module. Rather than operating as two isolated branches, these components form a tightly coupled system designed for semantic-guided and adaptive de-raining.

Specifically, the VAE encoder processes the input rainy image I_{in} and encodes it into a compact latent vector z, which

captures global rain-related semantics. This latent representation not only regularizes learning via coarse reconstruction but also serves as a semantic prior injected into each Transformer stage through Latent-Guided Attention (LGA) modules. Formally, the VAE models the posterior distribution as:

$$z \sim q_\phi(z|I_{\text{in}}) = \mathcal{N}(\mu_\phi(I_{\text{in}}), \sigma^2_\phi(I_{\text{in}})), \tag{1}$$

where q_ϕ is learned via variational inference [8].

Concurrently, the LGTformer extracts multi-scale spatial features through a stack of Transformer blocks, where LGA modules dynamically inject the latent code z to guide feature refinement. This latent-conditioned attention enables the model to disambiguate rain streaks from background structures more effectively.

B. Latent Prior Extraction via VAE

The VAE module extracts global semantic priors that guide the de-raining process. It follows a standard encoder-decoder architecture, where the encoder E_{vae} encodes the input rainy image I_{in} into a latent distribution $q_\phi(z|I_{\text{in}})$, assumed to be Gaussian. The decoder D_{vae} reconstructs a coarse clean image $\hat{I}_{\text{vae}}$ from a sampled latent code $z \sim q_\phi(z|I_{\text{in}})$, providing an auxiliary supervision signal and encouraging meaningful latent representations.

The VAE is trained by minimizing a variational lower bound, combining the Kullback-Leibler divergence between the approximate posterior and the prior $p(z)$, as well as an L_1 reconstruction loss between the decoder output and the ground truth clean image I_{gt}:

$$\mathcal{L}_{\text{VAE}} = D_{\text{KL}}(q_\phi(z|I_{\text{in}}) \parallel p(z)) + \|D_{\text{vae}}(z) - I_{\text{gt}}\|_1. \tag{2}$$

This learning objective encourages the latent space to capture global rain-relevant semantics, which can be injected into the restoration backbone via the LGA modules to facilitate structure-aware feature modulation.

C. Hierarchical Transformer Backbone

The LGTformer backbone employs a sequence of Transformer blocks enhanced with local-global reasoning modules [21]. It takes both spatial features and latent code z as input. Each block contains a Latent-Guided Attention module (LGA), a Mixed Attention Feed-forward Network (MAFN), and a Dual-Branch Mixed Fusion (DBMF) layer.

1) Latent-Guided Attention (LGA): This module performs cross-attention between the latent prior z and current spatial features x. Let $x \in \mathbb{R}^{H\times W\times C}$ and $z \in \mathbb{R}^C$, where C denotes the number of feature channels. The attention operation [21] is defined as:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{3}$$

where $Q = W_q x$, $K = W_k z$, and $V = W_v z$.

2) Mixed Attention Feed-Forward Network (MAFN): The MAFN module fuses grouped convolutions with kernel sizes 1×1, 3×3, and 5×5, followed by depthwise convolution and aggregation. The final fusion output is computed as [22]:

$$F_{\text{MAFN}} = \text{Conv}_{1\times1}\left([\text{Conv}_{1\times1}(x); \text{Conv}_{3\times3}(x); \text{Conv}_{5\times5}(x)]\right). \tag{4}$$

3) Dual-Branch Mixed Fusion (DBMF): DBMF merges shallow and deep features using parallel convolutional branches and channel fusion [23]:

$$F_{\text{DBMF}} = \text{Conv}_{1\times1}([F_1; F_2]), \tag{5}$$

where F_1 and F_2 represent outputs from branches with different resolutions.

D. Mixture-of-Experts (MoE) Routing

In real-world scenes, rain exhibits significant spatial heterogeneity, including variations in density, orientation and scale, as well as occlusion across different regions of an image. To address this inherent diversity, we incorporate Mixture-of-Experts (MoE) modules into both the semantic prior pathway (VAE branch) and the spatial restoration backbone (LGTformer), as illustrated in Fig. 1. The objective is to dynamically activate specialized feature-processing subnetworks that can better adapt to local content characteristics.

Each MoE block consists of N parallel expert networks $\{\text{Expert}_i\}_{i=1}^N$ and a lightweight gating MLP (MLP_g), which takes a shared input feature $x \in \mathbb{R}^C$ and produces normalized attention weights $\{\omega_i\}_{i=1}^N$ across the experts. The gating mechanism follows a softmax distribution [24]:

$$\omega_i = \frac{\exp(\text{MLP}_g(x)_i)}{\sum_{j=1}^N \exp(\text{MLP}_g(x)_j)}. \tag{6}$$

Each expert network receives the same input x but processes it differently due to distinct learned parameters. The final MoE output is computed as the weighted sum of all expert outputs [24]:

$$F_{\text{MoE}} = \sum_{i=1}^N \omega_i \cdot \text{Expert}_i(x). \tag{7}$$

This design allows the model to route features selectively through expert paths that are most suitable for a given local context. For example, in areas with fine rain streaks, experts trained on high-frequency structures are prioritized. One the other hand, in regions with dense or overlapping rain, coarse-grained experts dominate the output.

In the VAE branch, MoE enhances the flexibility of latent prior modeling by selectively decoding rain-invariant semantic components. In the spatial restoration backbone, MoE improves robustness by enabling context-aware feature refinement at different Transformer stages, particularly in scenes with strong spatial discontinuities or non-uniform degradation.

By explicitly modeling conditional specialization through MoE, our architecture gains finer control over feature routing,

TABLE I: Quantitative comparison of PSNR (dB) and SSIM on five benchmark datasets.

Method	Rain200L		Rain200H		DID-Data		DDN-Data		SPA-Data	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DSC [2]	27.16	0.8663	14.73	0.3815	24.24	0.8279	27.31	0.8373	34.95	0.9416
GMM [25]	28.66	0.8652	14.50	0.4164	25.81	0.8344	27.55	0.8479	34.30	0.9428
DDN [6]	34.68	0.9671	26.05	0.8056	30.97	0.9116	30.00	0.9041	36.16	0.9457
RESCAN [26]	36.09	0.9697	26.75	0.8353	33.38	0.9417	31.94	0.9345	38.11	0.9707
PReNet [13]	37.80	0.9814	29.04	0.8991	33.17	0.9481	32.60	0.9459	40.16	0.9816
MSPFN [27]	38.58	0.9827	29.36	0.9034	33.72	0.9550	32.99	0.9333	43.43	0.9843
RCDNet [15]	39.17	0.9885	30.24	0.9048	34.08	0.9532	33.04	0.9472	43.36	0.9831
MPRNet [16]	39.47	0.9825	30.67	0.9110	33.99	0.9590	33.10	0.9347	43.64	0.9844
DualGCN [17]	40.73	0.9886	31.15	0.9125	34.37	0.9620	33.01	0.9489	44.18	0.9902
SPDNet [28]	40.50	0.9875	31.28	0.9207	34.57	0.9560	33.15	0.9457	43.20	0.9871
Uformer [19]	40.20	0.9860	30.80	0.9105	35.02	0.9621	33.95	0.9545	46.13	0.9913
Restormer [9]	40.99	0.9890	32.00	0.9329	35.29	0.9641	34.20	0.9571	47.98	0.9921
IDT [5]	40.74	0.9884	32.10	0.9344	34.89	0.9623	33.84	0.9549	47.35	0.9930
DRSformer [20]	41.23	0.9894	32.17	0.9326	35.35	0.9646	34.35	0.9588	48.54	0.9924
VAEformer (Ours)	**41.12**	**0.9890**	**32.28**	**0.9331**	**35.40**	**0.9633**	**34.43**	**0.9582**	**48.60**	**0.9922**

reduces representation conflicts, and improves generalization across diverse rain scenarios. This is especially beneficial in scenes with mixed rain types, cluttered backgrounds, or low visibility, where fixed-path architectures often fail to respond adaptively.

E. Training Objective

The overall training objective supervises both the image-level restoration and the latent prior learning. Specifically, the final loss function combines a reconstruction loss $\mathcal{L}_{\text{img}}$ applied to the Transformer output and a variational loss $\mathcal{L}_{\text{VAE}}$ applied to the VAE branch, i.e., $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{img}} + \lambda \mathcal{L}_{\text{VAE}}$, where the reconstruction loss $\mathcal{L}_{\text{img}}$ is defined as the pixel-wise L_1 distance between the predicted clean image and the ground truth: $\mathcal{L}_{\text{img}} = \|I_{\text{pred}} - I_{\text{gt}}\|_1$. The hyperparameter λ balances the two terms and is empirically set to 0.1 in all experiments to ensure stable optimization and meaningful latent encoding.

This dual-objective training ensures that the network not only learns accurate image restoration, but also benefits from structured latent priors that improve generalization across diverse rain scenarios.

IV. Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our proposed VAE-Enhanced Transformer using Rain200L/H [31], DID-Data [32], DDN-Data [29], and SPA-Data [30]. These datasets contain both synthetic and real-world rain conditions. Our method is compared against prior-based, CNN-based, and Transformer-based state-of-the-art approaches on five widely-used benchmark datasets. For quantitative evaluation, we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR evaluates the fidelity at pixel-level restoration, while SSIM evaluates similarity in luminance, contrast, and structure. In addition, qualitative visualization is also employed for benchmarking purposes.

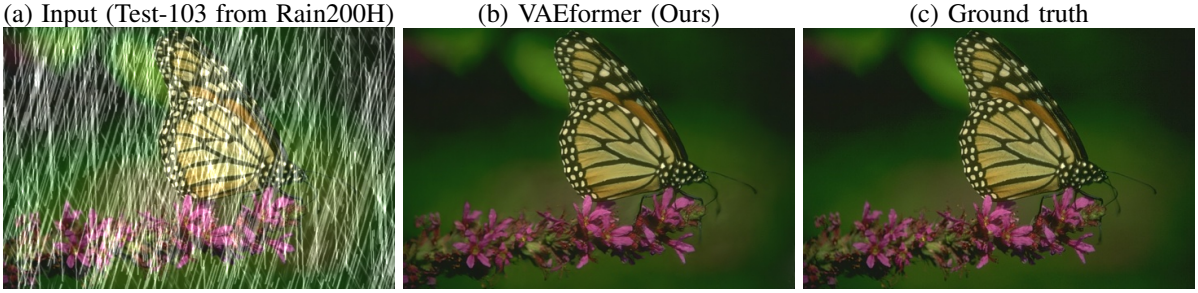
A. Comparison with State-of-the-Art Methods

We compare our model against three categories of baselines: (1) traditional prior-based methods, i.e., DSC by Luo et al. [2] and GMM by Li et al. [25]); (2) CNN-based approaches, i.e., DDN by Fu et al. [6], RESCAN by Li et al. [26], RCDNet by Wang et al. [15]), and; (3) recent Transformer-based architectures, i.e., Uformer by Wang et al. [19], Restormer by Zamir et al. [9], and DRSformer by Chen et al. [20]. Table I summarizes the PSNR and SSIM scores across all datasets, providing a comprehensive view of each method's restoration fidelity and perceptual quality.

Results suggest that **VAEformer** consistently outperforms all competing methods across five benchmark datasets, achieving the highest PSNR and SSIM scores in every case. When compared to leading Transformer-based baselines, VAEformer delivers notable improvements. For example, PSNR gains of +0.28dB over Restormer and +0.11dB over DRSformer are observed. These margins are particularly prominent on challenging datasets such as Rain200H and DID-Data, which feature dense, overlapping, and structurally entangled rain streaks. Such results reflect our model's ability to generalize under severe degradation conditions.

Beyond quantitative metrics, visual comparisons further confirm the perceptual advantages of VAEformer. Our model not only eliminates rain streaks more effectively but also maintains sharper edges, richer textures, and more consistent scene structure than conventional methods. This achievement arises from three key architectural contributions: (1) VAE-based latent priors: The integration of variational priors provides semantically meaningful, distribution-aware global guidance. This enhances long-range coherence and disambiguates rain patterns from structural content during decoding; (2) Hierarchical feature fusion: Through the use of Multi-Axis Fusion Networks (MAFN) and Dual-Branch Multi-scale Fusion (DBMF), the model efficiently aggregates features across spatial resolutions and contexts, improving the recovery of

(a) Input (Test-103 from Rain200H) (b) VAEformer (Ours) (c) Ground truth



(a) Input (Test-152 from Rain200H) (b) VAEformer (Ours) (c) Ground truth

Fig. 2: Examples of input rain image (1st column), de-rained image produced by the proposed VAEformer (2nd column), and ground truth (3rd column). The corresponding PSNR (SSIM) values are 32.26 dB (0.9328) and 32.19 dB (0.9311), respectively.

high-frequency details and preserving geometric consistency; (3) Mixture-of-Experts (MoE) routing: The MoE mechanism dynamically selects expert sub-modules conditioned on local feature characteristics. This makes the model adaptively specializes in different rain intensities, streak patterns, and background textures, boosting its flexibility and robustness.

These innovations collectively endow VAEformer with better generalization capabilities across both synthetic and real-world rain scenarios, setting a new performance standard in single-image de-raining.

B. Qualitative Results

To further assess the feasibility of the proposed VAEformer model in de-raining, we consider the perceptual quality of its de-rained output. Three representative images from different benchmark datasets are shown in Fig. 2 for visual inspection. Here, the original, output, and ground truth images are shown on the first, second, and third columns, respectively. These visualizations are particularly important for capturing perceptual factors that PSNR and SSIM may overlook, such as edge continuity, fine-grained textures, and the suppression of visual artifacts. They also reflect the model's generalization capability across varied spatial layouts, lighting conditions, and rain densities.

Specifically, for the input image on the first row, VAEformer successfully removes thin streaks while recovering clear building contours and lane markings that are partially occluded in the rainy input. For the input image on the second row, note that there are diagonal and overlapping streaks under uneven lighting, which are successfully removed by VAEformer. The results highlight our model's ability to restore consistent illumination and eliminate compounded degradations.

These results demonstrate the model's robustness in handling spatially diverse rain patterns and its capacity to maintain perceptual realism. The joint use of VAE-guided latent priors, latent-conditioned attention (LGA), and Mixture-of-Experts (MoE) enables semantic-aware restoration with adaptive specialization.

C. Ablation Test

To investigate the contribution of each module in our framework, we conduct an ablation study on the Rain200H dataset and summarized the results in Table II. Here, we use M1 (DRSformer [20]) as the baseline. Component A refers to the VAE-based rain representation with simple weighted summation, B denotes the expert-based fusion strategy, and C corresponds to the local-global attention module. From the table, we observe that each individual component contributes positively to the overall performance. Using only component A (M2) leads to limited gains, while using only B (M3) shows better results. Combining A and B (M4) further improves both PSNR and SSIM. Finally, incorporating all three components (A $\oplus$ B $\oplus$ C) achieves the best performance, surpassing the baseline (M1) and demonstrating the complementary nature of each module.

Together, the three components enable our model to effectively capture complex rain patterns and scene details. Similar performance trends are also observed on other datasets (e.g., Rain800 and Rain1400), and hence the discussions are omitted here for brevity.

V. Conclusion

In this work, we propose a novel VAE-enhanced Transformer framework for single-image de-raining, which integrates global latent priors and hierarchical attention to effectively disentangle rain streaks from structural scene

TABLE II: Ablation study on Rain200H. Evaluation is conducted on PSNR and SSIM.

Methods	A	B	C	PSNR↑	SSIM↑
M1				32.17	0.9326
M2	✓			29.65	0.8933
M3		✓		30.79	0.9112
M4	✓	✓		31.25	0.9262
Our Method	✓	✓	✓	**32.28**	**0.9331**

content. By leveraging the expressive power of a variational autoencoder, our method captures implicit rain patterns in the latent space and guides the feature learning process through latent-guided attention. The proposed dual-branch fusion architecture—comprising semantic-aware and spatial-aware processing paths—facilitates cross-scale interaction between contextual priors and high-resolution details, enabling fine-grained structure preservation without over-smoothing. Our work advocates a principled fusion of generative modeling and discriminative restoration, highlighting the benefits of integrating structured priors into Transformer architectures. Experiments on five benchmarks demonstrate that our model consistently outperforms existing CNN- and Transformer-based approaches in both quantitative metrics and visual restoration fidelity.

As future work, we plan to extend this paradigm to multi-modal, weather-aware restoration and explore its deployment in video and real-time de-raining systems.

REFERENCES

[1] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D*, **60**(1–4), 259–268 (1992)

[2] Luo, Y., Xu, Y., Ji, H.: Removing rain from a single image via discriminative sparse coding. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV), pp. 3397–3405 (2015)

[3] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using Swin Transformer. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 1833–1844 (2021)

[4] Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Single image rain streak decomposition using layer priors. *IEEE Trans. Image Process.* **26**(8), 3874–3885 (2017)

[5] Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.-J.: Image de-raining transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(11), 12978–12995 (2022)

[6] Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3855–3863 (2017)

[7] Tan, F., Kong, Y., Fan, Y., Liu, F., Zhou, D., Chen, L., Gao, L., Qian, Y.: SDNet: Multi-branch for single image deraining using Swin. arXiv preprint arXiv:2105.15077 (2021)

[8] Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)

[9] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5728–5739 (2022)

[10] Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C.: EDVR: Video restoration with enhanced deformable convolutional networks. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 0–0 (2019)

[11] Zhang, R., Cahyawijaya, S., Cruz, J.C.B., Winata, G.I., Aji, A.F.: Multilingual large language models are not (yet) code-switchers. arXiv preprint arXiv:2305.14235 (2023)

[12] He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(12), 2341–2353 (2010)

[13] Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: A better and simpler baseline. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3937–3946 (2019)

[14] Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 695–704 (2018)

[15] Wang, H., Xie, Q., Zhao, Q., Meng, D.: A model-driven deep neural network for single image rain removal. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3103–3112 (2020)

[16] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Multi-stage progressive image restoration. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 14821–14831 (2021)

[17] Fu, X., Qi, Q., Zha, Z.-J., Zhu, Y., Ding, X.: Rain streak removal via dual graph convolutional network. In: Proc. AAAI Conf. on Artificial Intelligence (AAAI), **35**(2), 1352–1360 (2021)

[18] Yang, W., Tan, R.T., Wang, S., Fang, Y., Liu, J.: Single image deraining: From model-based to data-driven and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(11), 4059–4077 (2020)

[19] Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 17683–17693 (2022)

[20] Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5896–5905 (2023)

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. **30** (2017)

[22] Howard, A.G.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[23] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 10012–10022 (2021)

[24] Fedus, W., Zoph, B., Shazeer, N.: Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **23**(120), 1–39 (2022)

[25] Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Rain streak removal using layer priors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2736–2744 (2016)

[26] Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: Proc. European Conf. on Computer Vision (ECCV), pp. 254–269 (2018)

[27] Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., Jiang, J.: Multi-scale progressive fusion network for single image deraining. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 8346–8355 (2020)

[28] Yi, Q., Li, J., Dai, Q., Fang, F., Zhang, G., Zeng, T.: Structure-preserving deraining with residue channel prior guidance. In: Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 4238–4247 (2021)

[29] Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3855–3863 (2017)

[30] Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.H.: Spatial attentive single-image deraining with a high quality real rain dataset. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 12270–12279 (2019)

[31] Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1357–1366 (2017)

[32] Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 695–704 (2018)