Semantics and Content Matter: Towards Multi-Prior Hierarchical Mamba for Image Deraining

Zhaocheng Yu[1, 2], Kui Jiang[1, 2, *], Junjun Jiang[1], Xianming Liu[1], Guanglu Sun[3], Yi Xiao[4]

[1]Faculty of Computing, Harbin Institute of Technology,
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ),
[3]School of Computer Science and Technology, Harbin University of Science and Technology,
[4]School of Computer and Artificial Intelligence, Zhengzhou University
yuzhaocheng@stu.hit.edu.cn, {jiangkui,jiangjunjun,csxm}@hit.edu.cn, Sunguanglu@hrbust.edu.cn, yixiao@zzu.edu.cn

Abstract

Rain significantly degrades the performance of computer vision systems, particularly in applications like autonomous driving and video surveillance. While existing deraining methods have made considerable progress, they often struggle with fidelity of semantic and spatial details. To address these limitations, we propose the Multi-Prior Hierarchical Mamba (MPHM) network for image deraining. This novel architecture synergistically integrates macro-semantic textual priors (CLIP) for task-level semantic guidance and micro-structural visual priors (DINOv2) for scene-aware structural information. To alleviate potential conflicts between heterogeneous priors, we devise a progressive Priors Fusion Injection (PFI) that strategically injects complementary cues at different decoder levels. Meanwhile, we equip the backbone network with an elaborate Hierarchical Mamba Module (HMM) to facilitate robust feature representation, featuring a Fourier-enhanced dual-path design that concurrently addresses global context modeling and local detail recovery. Comprehensive experiments demonstrate MPHM's state-of-the-art performance, achieving a 0.57 dB PSNR gain on the Rain200H dataset while delivering superior generalization on real-world rainy scenarios.

Introduction

Adverse weather conditions degrade image quality, impairing the performance of advanced computer vision algorithms. This poses significant challenges in critical applications, such as autonomous vehicles and surveillance systems (Dang et al. 2024, 2023; Hong et al. 2025).

The image deraining task aims to restore high-quality visual content from rain-affected inputs and is a fundamental component of modern computer vision systems. Traditional methods (Kang, Lin, and Fu 2011; Chen and Hsu 2013) use rain pattern analysis and handcrafted priors for rain removal. However, their performance is often constrained by strong scene-specific dependencies, resulting in poor generalization across diverse real-world scenarios (Zhong et al. 2022).

Recent advances in deep learning (LeCun, Bengio, and Hinton 2015) have significantly improved image deraining.

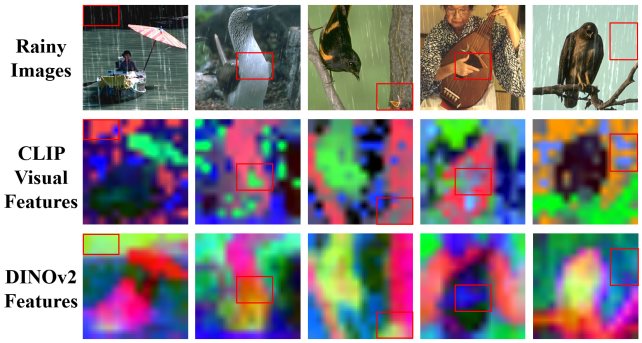

Figure 1: For analytical purposes, this visualization displays the PCA (Principal Component Analysis) projection of high-dimensional features from the visual encoders of various pre-trained models into 3-dimensional space.

CNN-based methods (Wang et al. 2020a; Guo et al. 2023; Ren et al. 2019) excel at recovering local textures through hierarchical convolutions but are constrained by limited receptive fields, hindering their ability to capture long-range dependencies. Transformer-based (Xiao et al. 2022; Chen et al. 2021) and Mamba-based (Li et al. 2024) approaches overcome this limitation by aggregating global response with self-attention or state space module (SSM), leading to substantial performance gain. However, these methods struggle with either insufficient semantic completeness or limited spatial detail preservation.

Recent breakthroughs in large models have generated significant interest in their scene perception capabilities (Radford et al. 2021; Oquab et al. 2023). These models typically employ an architecture where images are converted by a visual encoder into a series of visual tokens or by a text encoder into language-aligned feature descriptions (Bai et al. 2025). Inspired by this, large-model priors have also been applied to image restoration tasks, such as extracting the macro-level degradation description (Jin et al. 2024) or micro-level background textures (Lin et al. 2023), yielding substantial performance gains. Image deraining tasks encompass various rain patterns with distinct characteristics. This diversity makes it challenging for a single prior—whether textual or visual—to optimally balance se-

*Corresponding Author

mantic completeness and content fidelity across all scenarios. To address this, (Luo et al. 2023) integrate CLIP's textual and visual encoders (Radford et al. 2021) into a unified framework to provide comprehensive semantic guidance for image restoration. However, CLIP's visual encoder prioritizes global image-text alignment, favoring abstract semantics over detail. As shown in Figure 1, it is also sensitive to noise, producing suboptimal features for rainy scenes. This phenomenon may be due to CLIP's training data and objective, where the global-level image-text contrastive learning makes the image encoder poor at capturing visual details. In contrast, the DINOv2 encoder (Oquab et al. 2023) generates finer-grained representations with richer semantic content. This suggests a promising approach: harmonizing DINOv2's micro visual priors with CLIP's macro textual priors to adapt to diverse visual characteristics and requirements. Unfortunately, the independent architectures of CLIP and DINOv2 prevent joint optimization with multimodal data. Consequently, each encoder remains confined to its original pre-trained capabilities, limiting their potential to leverage synergistic advantages in multimodal representation.

To this end, we develop the Multi-Prior Hierarchical Mamba (MPHM) framework to complement semantics and details while alleviating modality conflicts encountered by multiple encoders for high-accuracy image deraining. To promote semantic completeness, MPHM sequentially integrates the scene-aware visual priors (micro) generated by DINOv2 and the task-level textual priors (macro) from CLIP into all decoder stages of the backbone network. This sequential Priors Fusion Injection (PFI) strategy significantly alleviates feature conflicts that may arise from direct fusion of multi-prior representations, while ensuring comprehensive semantic guidance throughout the network. To enhance texture representation, we introduce a Hierarchical Mamba Module (HMM) that enables multi-level global-local interactions and cross-channel knowledge integration. This design effectively addresses Mamba's limitations in capturing local spatial relationships. Furthermore, through combined spatial and frequency domain processing, HMM establishes robust structural and textural representations. In general, the contributions of this work can be summarized as follows.

- We propose a novel Multi-Prior Hierarchical Mamba (MPHM) framework that systematically integrates complementary semantic and detail priors from multiple pre-trained foundation models to achieve robust image deraining.
- We introduce a Priors Fusion Injection (PFI) where DINOv2's visual priors and CLIP's textual priors are progressively fused across decoder levels to jointly enhance spatial detail and semantic representation. A Hierarchical Mamba Module (HMM) is proposed to augment Mamba's local spatial modeling capability through multi-level global-local feature interactions, thereby improving structural representation and restoration robustness via integrated spatial-frequency processing.
- Extensive experiments on both synthetic and real-world datasets demonstrate that our proposed MPHM method consistently outperforms state-of-the-art approaches in terms of both quantitative metrics and qualitative results.

Method

This section details the proposed framework's workflow, optimization objectives, and key components.

Pipeline Overview

Figure 2 illustrates the MPHM architecture, comprising two core components: a U-shaped encoder-decoder backbone built with Hierarchical Mamba Modules (HMM), and a prior generation-injection pipeline. The backbone processes the rainy input I_{rain} to extract and refine dual-domain (spatial/frequency) features, with skip connections ensuring structural consistency. The prior generation system provides complementary semantic guidance. We generate the scene-level visual prior P_v from I_{rain} using a frozen DINOv2 encoder and lightweight adapter, while deriving the task-level textual prior P_t from handcrafted description ("No rain") via the frozen CLIP text encoder and refinement adapter. These priors are integrated at each decoder stage via Priors Fusion Injection (PFI) modules. PFI employs a dual-stage cross-attention mechanism to inject texture and task priors into the backbone features, followed by self-attention and a gated feedforward block (GDFN) for semantic propagation and refinement. The final derained image I_{pred} is reconstructed by eliminating the predicted rain residual from I_{rain}.

Model Optimization. For faithful deraining and texture preservation, we minimize a hybrid loss that integrates pixel reconstruction with frequency-domain contrastive learning. The reconstruction loss ensures spatial accuracy:

$$\mathcal{L}_{rec} = \sum_{i=1}^{n} \left\| I_{pred} - I_{gt} \right\|_1 , \tag{1}$$

where I_{gt} and I_{pred} denote the ground-truth and predicted rain-free images respectively. To enhance frequency-domain consistency, we incorporate the Frequency-domain Contrastive Regularization (FCR) loss (Gao et al. 2024) that minimizes the distance between derained spectra and clean references while maximizing their separation from rainy patterns. The FCR loss is formulated as:

$$\mathcal{L}_{fcr} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left\| \mathcal{F}(I_{gt}) - \mathcal{F}(I_{pred}) \right\|_1}{\left\| \mathcal{F}(I_{rain*}) - \mathcal{F}(I_{pred}) \right\|_1}, \tag{2}$$

where derained spectra $\mathcal{F}(I_{pred})$ are anchors, clean spectra $\mathcal{F}(I_{gt})$ are positives, and randomly sampled rainy spectra $\mathcal{F}(I_{rain*})$ are negatives. $\mathcal{F}(\cdot)$ operator denotes the Discrete Fourier Transform (DFT) and parameter $n = 2$ indicates the number of negative samples. The overall loss function is defined as:

$$\mathcal{L}_{Total} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{fcr}, \tag{3}$$

where $\lambda = 0.1$ is an empirically determined balancing weight that adjusts the trade-off between spatial-domain accuracy and frequency-domain fidelity.

Hierarchical Mamba Module

As shown in Figure 3, HMM processes features via parallel spatial and frequency branches.

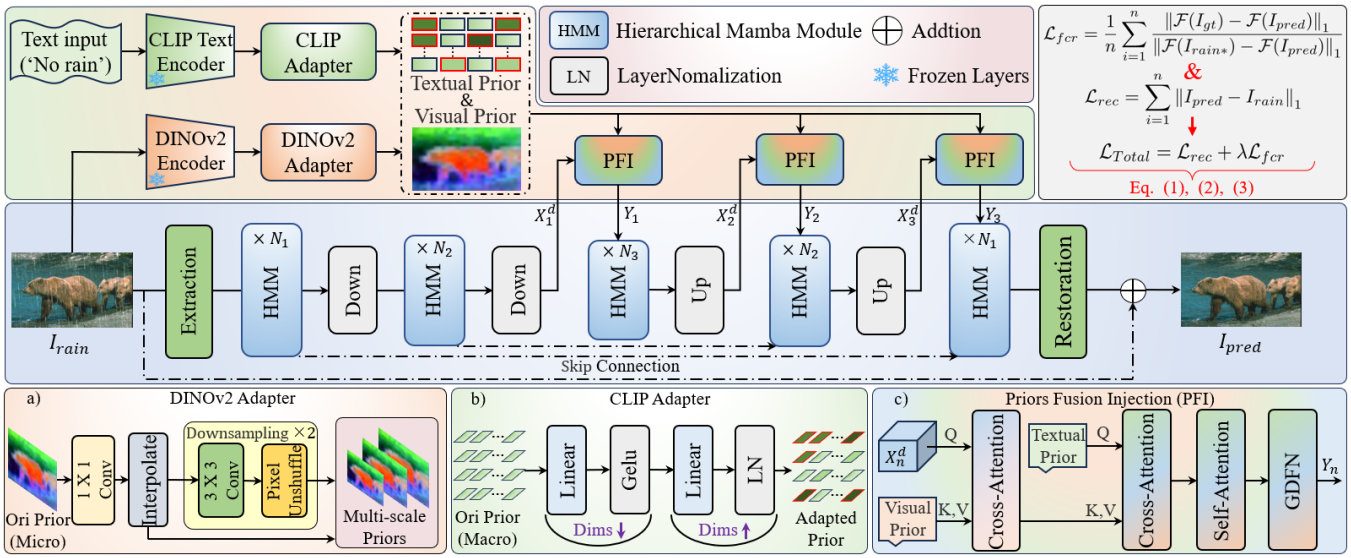

Figure 2: Overview of the proposed Multi-Prior Hierarchical Mamba deraining framework. (a) DINOv2 Adapter, (b) CLIP Adapter and (c) Priors Fusion Injection (PFI).

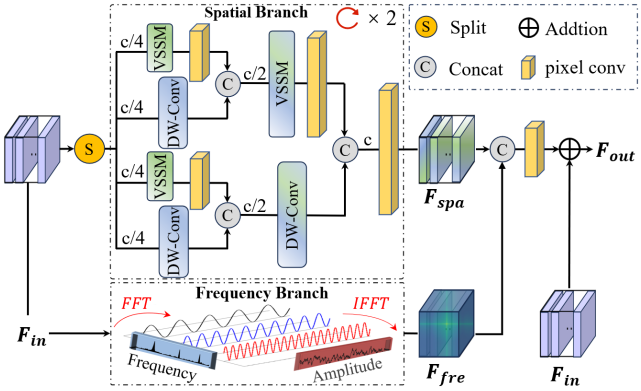

Figure 3: Pipeline of the HMM.

Spatial-domain Branch. To enhance local spatial modeling in Mamba-based architectures, we introduce a spatial branch that hierarchically integrates global and local features. Given input $F_{in} \in \mathbb{R}^{(b,c,h,w)}$, we first split features channel-wise into four equal partitions: $[f_1, f_2, f_3, f_4] = \mathcal{S}(F_{\text{in}})$ with $f_i \in \mathbb{R}^{(b,\frac{c}{4},h,w)}$. These sub-features undergo parallel processing through alternating modules. Specifically, f_1 and f_3 are passed through Visual Selective Spatial Mamba (VSSM) (Liu et al. 2024) for global context modeling, while f_2 and f_4 are processed by depth-wise convolution (DW-Conv) to improve local feature continuity. VSSM outputs are refined via 1×1 convolutions maintaining spatial resolution. We then concatenate the processed features pairwise ((f_1, f_2) and (f_3, f_4)) to form two intermediate features, denoted as F_{GL1} and F_{GL2}. To strengthen global-local coupling, we further process these intermediates, where F_{GL1} is first processed by a VSSM block and a 1×1 convolution while F_{GL2} is passed through a DW-Conv. The refined intermediate features are concatenated and passed through a 1×1 convolution to generate final spatial representation F_{Spa}. This hierarchical design enjoys these merits: *i*) multi-stage fusion of global context and local details; *ii*) concurrent capturing of fine-grained textures and holistic spatial patterns; *iii*) progressively promoting the cross-context features through iterative refinement.

Frequency-domain Branch. Following (Gao et al. 2024), our frequency branch employs a 2D Fast Fourier Transform (FFT)-based architecture called FFCM. This extracts frequency-domain features F_{fre} from the input scene. By combining point-wise and multi-scale depth-wise convolutions, FFCM enhances its representational capacity while maintaining computational efficiency.

Dual-domain Fusion. To integrate the outputs of both the spatial and frequency branches, we concatenate them and apply a 1×1 convolution for channel compression. This fused representation is then added to the original input F_{in} in a residual manner, producing the final output F_{out}. This design enables comprehensive scene feature extraction while maintaining computational efficiency.

Multi-modal Priors Guidance

Effective multi-prior guidance requires resolving inter-level feature conflicts. As shown in Figure 2, we achieve this via two components: *i*) domain-specific feature adapters aligning prior-backbone representational spaces, and *ii*) the Priors Fusion Injection (PFI) module alleviating conflicts during integration via structured attention.

DINOv2 Adapter. The DINOv2 prior P_v undergoes a three-stage adaptation for backbone integration: *i*) initial channel reduction via pixel-wise convolution; *ii*) spatial alignment via bilinear interpolation to match backbone resolutions; *iii*) hierarchical representation generation using convolutional layers with PixelUnshuffle downsampling. This ensures dimensionally compatible prior injection across decoder stages while preserving structural semantics.

Type	Methods	Rain200H		Rain200L		DID-Data		DDN-Data		SPA-Data		Average	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Prior	DSC (Luo, Xu, and Ji 2015)	14.73	0.3815	27.16	0.8663	24.24	0.8279	27.31	0.8373	34.95	0.9416	25.67	0.7709
	GMM (Li et al. 2016)	14.50	0.4164	28.66	0.8652	25.81	0.8344	27.55	0.8479	34.30	0.9428	26.16	0.7813
CNN	RESCAN (Li et al. 2018)	26.75	0.8353	36.09	0.9697	33.38	0.9417	31.94	0.9345	38.11	0.9707	33.25	0.9303
	PReNet (Ren et al. 2019)	29.04	0.8991	37.80	0.9814	33.17	0.9481	32.60	0.9459	40.16	0.9816	34.55	0.9512
	MSPFN (Jiang et al. 2020)	29.36	0.9034	38.58	0.9827	33.72	0.9550	32.99	0.9333	43.43	0.9843	35.61	0.9517
	RCDNet (Wang et al. 2020b)	30.24	0.9048	39.17	0.9885	34.08	0.9532	33.04	0.9472	43.36	0.9831	35.97	0.9553
	MPRNet (Zamir et al. 2021)	30.67	0.9110	39.47	0.9825	33.99	0.9590	33.10	0.9347	43.64	0.9844	36.17	0.9543
	DualGCN (Fu et al. 2021)	31.15	0.9125	40.73	0.9886	34.37	0.9620	33.01	0.9489	44.18	0.9902	36.68	0.9604
	SPDNet (Yi et al. 2021)	31.28	0.9207	40.50	0.9875	34.57	0.9560	33.15	0.9457	43.20	0.9871	36.54	0.9594
	FADfomer (Gao et al. 2024)	32.48	0.9359	41.80	0.9906	35.48	0.9657	34.42	0.9602	49.21	0.9934	38.67	0.9691
Transformer	Uformer (Wang et al. 2022)	30.80	0.9105	40.20	0.9860	35.02	0.9621	33.95	0.9545	46.13	0.9913	37.22	0.9608
	Restormer (Zamir et al. 2022)	32.00	0.9329	40.99	0.9890	35.29	0.9641	34.20	0.9571	47.98	0.9921	38.09	0.9670
	IDT (Xiao et al. 2022)	32.10	0.9344	40.74	0.9884	34.89	0.9623	33.84	0.9549	47.35	0.9930	37.78	0.9666
	HCT-FFN (Chen et al. 2023c)	31.51	0.9100	39.70	0.9850	33.96	0.9592	33.00	0.9502	45.79	0.9898	36.79	0.9588
	DRSformer (Chen et al. 2023a)	32.17	0.9326	41.23	0.9894	35.35	0.9646	34.35	0.9588	48.54	0.9924	38.32	0.9675
	MSDT (Chen et al. 2024)	32.45	0.9379	41.75	0.9904	35.37	0.9652	34.36	0.9593	49.07	0.9926	38.60	0.9690
	NeRD-Rain (Chen, Pan, and Dong 2024)	32.40	0.9373	41.71	0.9903	35.53	0.9659	34.45	0.9596	49.58	0.9940	38.73	0.9694
Mamba	TransMamba (Sun et al. 2024)	32.96	0.9409	41.92	**0.9938**	35.63	0.9657	34.72	0.9603	49.72	**0.9968**	38.99	0.9715
	FourierMamba (Li et al. 2024)	32.71	0.9395	42.27	0.9908	35.49	0.9659	**35.58**	0.9599	49.18	0.9931	39.04	0.9698
	MPHM(Ours)	**33.53**	**0.9475**	**42.30**	0.9913	**35.65**	**0.9662**	34.80	**0.9607**	**49.76**	0.9951	**39.21**	**0.9722**

Table 1: Quantitative comparison of PSNR/SSIM on five benchmark test datasets. Our MPHM achieves the best or comparable performance across all benchmarks. The best result for each dataset is highlighted in bold, and the second-best is underlined.

CLIP Adapter. A lightweight bottleneck adapter processes the CLIP text prior: *i*) using linear projection to reduce dimensionality to a latent space; *ii*) employing nonlinear layers to extract task-specific semantics; *iii*) aligning features to visual representations.

Priors Fusion Injection. The adapted priors are integrated through a two-stage attention mechanism. Using DINOv2's scene-aware information P_v as keys/values provides cues for structural semantic retrieval. Employing CLIP-derived features P_t as queries focuses on task-relevant regions. This sequential strategy prevents fusion conflicts. A subsequent multi-head self-attention module captures long-range dependencies. Finally, a Gated Depth-wise Feedforward Network (GDFN) (Zamir et al. 2022) refines the features via its dual-gating mechanism, balancing transformation and spatial preservation.

Experiments

We detail MPHM's implementation and training protocols and then comprehensively evaluate its effectiveness. We compare against 19 state-of-the-art deraining methods over the last decade on synthetic and real-world datasets with multiple metrics.

Implementation Details

Datasets. The proposed method is evaluated on five benchmark datasets: Rain200L/H (Yang et al. 2017), DID-Data (Zhang and Patel 2018), DDN-Data (Fu et al. 2017), and SPA-Data (Wang et al. 2019), involving various rain patterns and scene complexities. The Rain200L dataset comprises 1,800 synthetically generated rainy/clean image pairs for training, along with a separate test set of 200 image pairs. As its counterpart, Rain200H maintains identical data partitioning while exclusively containing heavy rain conditions. The DID-Data provides 12,000 synthetic training pairs accompanied by 1,200 dedicated test pairs, whereas DDN-Data contains 12,600 training pairs with 1,400 test samples. SPA-Data is a large-scale real-world benchmark that contains 638,492 training pairs and a standardized test set of 1,000 image pairs with ground truth. Furthermore, we evaluate the model's generalization capability on two real-world datasets: Internet-Data (Wang et al. 2019) and RE-RAIN (Chen et al. 2023b), containing 147 and 300 rainy images without ground truth, respectively.

Metrics. For benchmark datasets with ground truth, we report PSNR (Huynh-Thu and Ghanbari 2008) and SSIM (Wang et al. 2004) for evaluation. For unpaired real-world data, we use no-reference metrics: BRISQUE (Mittal, Soundararajan, and Bovik 2012) and NIQE (Mittal, Moorthy, and Bovik 2012) to evaluate perceptual quality.

Experimental Settings. The HMM depth per stage is $\{4, 6, 8, 6, 4\}$ with an initial channel dimension $C = 32$. Model training employs the Adam optimizer (Loshchilov, Hutter et al. 2017) with a 1×10^{-3} initial learning rate decayed to 1×10^{-5} via cosine annealing (Loshchilov and Hutter 2016). The inputs are center-cropped to 256×256 resolution with a batch size of 4. All experiments run on PyTorch (Paszke 2019) using NVIDIA 3090 GPUs.

Comparison with State-of-the-arts

Synthetic Datasets Results. Quantitative results in four synthetic datasets are shown in Table 1, demonstrating that our MPHM outperforms existing approaches with an obvious performance gain. In the Rain200H and Rain200L datasets, MPHM attains a cumulative improvement of 0.85 dB, showing superior generalization in degradation com-

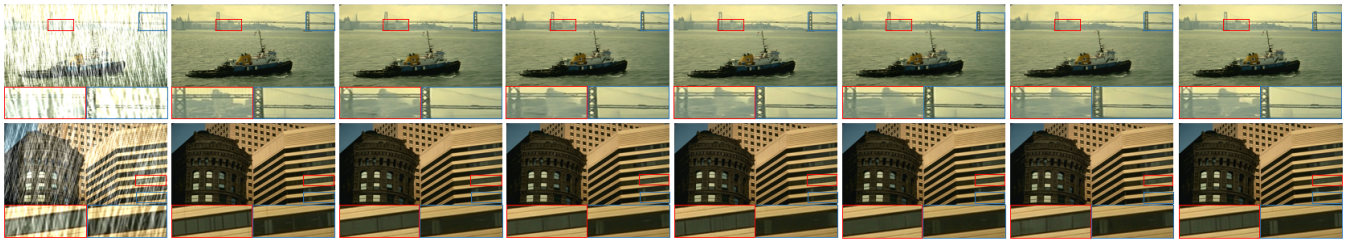

Figure 4: Visual results on Rain200H. Our method recovers clearer details and textures. Please zoom in for a better view.

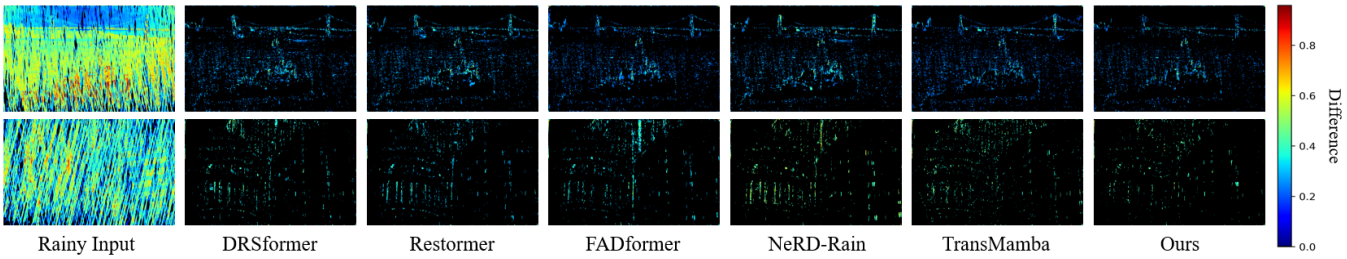

Figure 5: Residual heatmaps between derained results and ground truth. Brighter colors denote larger pixel-wise deviations. Our method shows minimal differences.

Methods	**Rainy Input**	**HCT-FFN**	**Restormer**	**IDT**	**DRSformer**	**NeRD-Rain**	**FADformer**	**TransMamba**	**Ours**
BRISQUE ↓	25.279	35.425	28.410	23.704	22.857	21.693	21.670	23.420	**21.221**
NIQE ↓	4.169	4.938	4.106	3.918	3.875	3.862	3.980	3.844	**3.787**

Table 2: Quantitative evaluations on the unpaired RE-RAIN dataset. We report BRISQUE (↓) and NIQE (↓), no-reference image quality metrics where lower values indicate better perceptual quality. Our method achieves the best scores.

plexities. Figure 4 highlights the ability of MPHM to enhance texture preservation, including scenarios when background patterns resemble rain streaks (*e.g.*, oblique or vertical lines). Figure 5 visualizes the pixel-wise residuals. It is obvious that our method yields sparser, lower-intensity residuals (darker heatmaps), confirming effective rain removal and detail preservation. This considerable superiority comes from semantic information supplementation and confusion elimination, where conventional methods mistakenly remove object structures alongside rain artifacts. Our semantic prior injection enables precise residual-background separation, maximizing detail retention in heavy rain.

Real-world Datasets Results. We further validate our approach on the real-world SPA-Data benchmark. As shown in Table 1, our MPHM achieves the best performance in all indicators. Figure 6 shows that the derained images produced by our MPHM enjoy better texture preservation and less rain residue. Table 2 tabulates the comparison of cross-domain generalization, where models trained on the Rain200H dataset are tested on unpaired RE-RAIN data. As expected, our method achieves the lowest BRISQUE and NIQE scores, indicating superior perceptual quality. Figure 7 further validates the robust capability of rain removal and detail reconstruction in real-world scenarios, bridging the synthetic-to-real domain gap.

Model Complexity. To evaluate the model efficiency, we measure parameters and FLOPs using 256×256 input. As summarized in Table 3, MPHM maintains a favorable computational profile while balancing deraining performance and inference efficiency.

Methods	Restormer	IDT	DRSformer	FADformer
FLOPs (G)	174.7	61.90	220.33	48.51
Params (M)	26.12	16.41	33.65	**6.96**
PSNR (db)	38.09	37.78	38.32	38.67
Methods	NeRD-Rain	TransMamba	FourierMamba	Ours
FLOPs (G)	148.05	45.67	**22.56**	61.89
Params (M)	22.89	16.74	17.62	10.28
PSNR (db)	38.73	38.99	39.04	**39.21**

Table 3: Comparison of model complexity and performance against state-of-the-art methods. The size of the test image is 256×256 pixels. "FLOPs" (in G) and "Params" (in M) denote the floating-point operations and the number of trainable parameters, respectively.

Ablation Studies

Ablation studies are conducted on the Rain200H with consistent training configurations to allow fair performance comparison between variants.

Effectiveness of HMM. As shown in Table 4, the results demonstrate the critical contributions of the HMM components. Removing the frequency-domain branch (FFCM) in Model-1 causes a 2.50 dB drop in PSNR, confirming

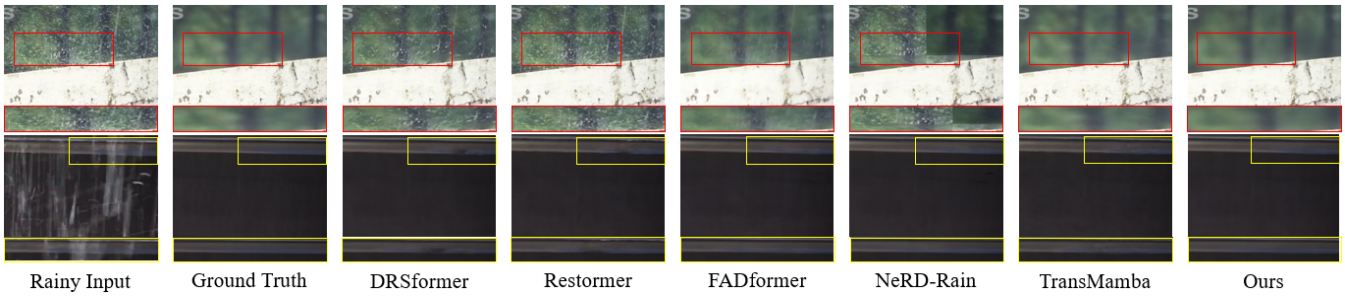

Figure 6: Derained results on the real-world paired SPA-Data. Compared with the derained results, our approach achieves superior rain removal performance and simultaneously preserves structural details. Please zoom in for a better view.

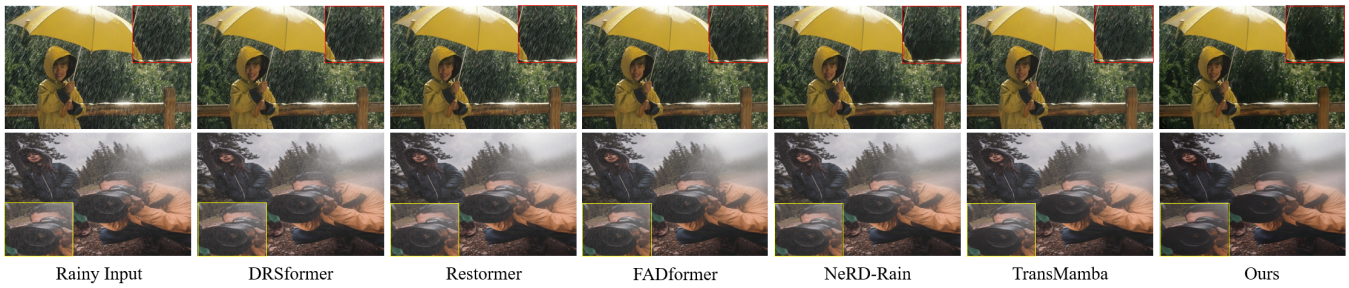

Figure 7: Visualizing on real-world unpaired RE-RAIN (top) and Internet-Data (bottom).

Variant	FFCM	DW	VSSM	PSNR	SSIM	Params	FLOPs
Model-1	×	✓	✓	30.56	0.9158	2.307	23.54
Model-2	✓	×	✓	32.93	0.9407	8.470	69.58
Ours	✓	✓	✓	33.06	0.9421	7.247	57.76

Table 4: Component-level ablations of the proposed HMM. The results validate the effectiveness of frequency and spatial encoding.

that frequency information is essential for preserving structural patterns and suppressing residual rain streaks. Model-2 shows a 0.13 dB PSNR reduction compared to the full HMM while requiring 16.9% more parameters and 20.5% higher FLOPs, indicating suboptimal efficiency. Figure 8 further reveals Model-2's diminished local spatial awareness (manifested as blurred high-frequency textures) due to the omitted feature grouping and DW-Conv. Conversely, our grouped local-global strategy combines parallel DW-Conv (local refinement) with VSSM (global context), reducing complexity while enhancing local modeling for superior efficiency and fidelity.

Ablation on Branch Fusion Scheme. We evaluate three fusion strategies for integrating frequency and spatial branches in the HMM module: direct addition, cross-attention, and our concatenation-based approach with a 1×1 convolution. As shown in Table 5, our method achieves the highest PSNR/SSIM with moderate complexity. Direct addition yields slightly inferior performance due to feature redundancy and interference. Cross-attention incurs the highest computational cost and the worst restoration quality, suggesting over-parameterization. Our lightweight concatenation method thus optimally balances restoration efficacy and efficiency.

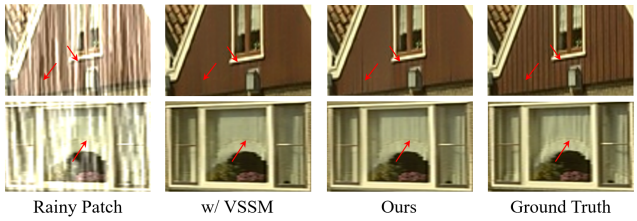

Figure 8: Qualitative comparison of local detail recovery. Our strategy enhances the VSSM's capacity to reconstruct spatial structures.

Strategy	Addition	Cross-attn	Concat (Ours)
PSNR/SSIM	32.82/0.9397	32.22/0.9339	33.06/0.9421
Params/FLOPs	6.864/53.74	7.352/ 59.31	7.247/57.76

Table 5: Ablation studies for branch fusion scheme.

Ablation on Priors Injection. We ablate four prior-guided representation learning configurations, involving *no priors*, *visual prior* (P_v) only, *textual prior* (P_t) only, and joint injection (Table 6). The baseline without priors achieves only 33.06 dB PSNR and 0.9421 SSIM, exhibiting noisy, entangled features that confuse rain streaks from scene contents (Figure 10). Introducing P_v (from DINOv2) moderately improves structural recovery but overemphasizes rain streaks that resemble scene patterns. Using P_t alone enhances rain suppression, but produces over-smooth textures. Joint injection of both priors maximizes perfor-

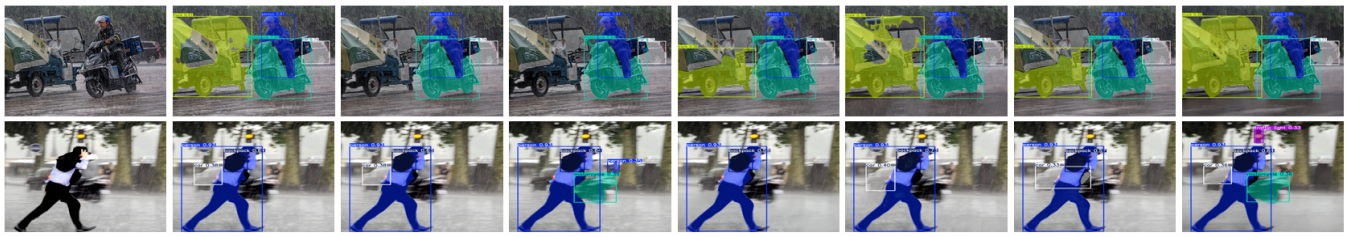

Figure 9: Visual comparison of object detection and instance segmentation enhanced by different deraining methods. Our method preserves clearer structure and yields more accurate visual perception. Please zoom in for a better view.

Variant	P_v	P_t	PSNR	SSIM
$w/o\ Priors$	×	×	33.06	0.9421
$w\ P_v$	✓	×	33.22	0.9446
$w\ P_t$	×	✓	33.20	0.9434
$w\ P_v \& \ P_t (Ours)$	✓	✓	33.53	0.9475

Table 6: Ablation studies on priors injection in PFI. "P_v" and "P_t" indicate the visual and textual priors, respectively.

mance by combining P_v's structural awareness with P_t's task-specific guidance, enabling robust semantic disentanglement and visually consistent deraining (Figure 10).

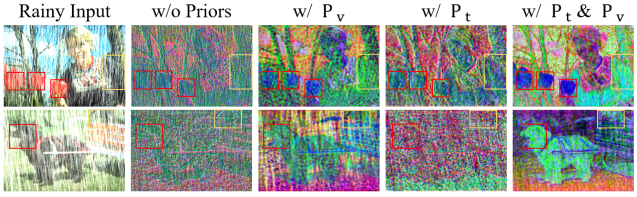

Figure 10: PCA visualization of different prior features from the bottleneck layer.

Ablation on Priors Fusion Scheme. We ablate four fusion strategies to integrate visual (P_v) and textual (P_t) priors, involving addition, concatenation, cross-attention, and our hierarchical fusion. As shown in Table 7, both the addition and concatenation schemes degrade restoration performance due to non-adaptive merging that equally weights features irrespective of semantic relevance. This propagates conflicting errors, making it difficult to learn semantic relation representations across modalities. Although the cross-attention scheme brings marginal gains, single-stage fusion forces incompatible merging of P_v's micro scene abstractions and P_t's macro task guidance. Our hierarchical fusion overcomes these limitations by decoupling injection: P_v adapts scene representations in early stages, while P_t refines task-specific features later. This progressive reconciliation eliminates semantic conflicts, enabling substantial semantic fusion and robust feature representation.

Impact on Downstream Task

To rigorously validate real-world utility, we evaluate derained images using YOLOv8 (Jocher 2023) for object detection and instance segmentation on the RE-RAIN dataset.

Strategy	Addition	Concat	Cross-attn	Ours
PSNR/SSIM	31.95/0.9287	32.05/0.9324	33.15/0. 9448	33.53/0.9475

Table 7: Ablation studies on priors fusion scheme.

As shown in Figure 9, our approach achieves substantial performance gains over state-of-the-art methods (better detection segmentation accuracy). These improvements directly stem from our dual technical innovations. i) The multi-prior fusion strategy (integrating DINOv2's pixel-level visual cues and CLIP's contextual language priors) enables adaptive scene representation, preserving structural coherence of objects under heavy occlusion. This resolves semantic ambiguity in degraded regions (*e.g.*, distinguishing rain-obscured cars from background clutter). ii) The Hierarchical Mamba Module (HMM) facilitates multi-scale feature refinement through global-local interactions in both spatial and frequency domains. By recovering high-frequency textures (critical edges and contours), HMM overcomes Mamba's inherent local detail limitations, enabling precise object boundary delineation. The synergy of semantic preservation and textural fidelity underpins MPHM's robustness for adverse-weather vision systems. Quantitative gains in detection/segmentation metrics confirm our method's superiority in maintaining task-critical information.

Conclusion

In this work, we proposed MPHM, a multi-prior hierarchical Mamba framework for single image deraining. By strategically combining macro-level textual priors from CLIP with micro-level visual priors from DINOv2, MPHM delivers comprehensive semantic guidance that improves discrimination between rain streaks and background structures. Furthermore, our Hierarchical Mamba Module (HMM) introduces a global-local interaction mechanism and Fourier-enhanced representations, effectively addressing the spatial locality limitations of conventional Mamba architectures. Extensive experiments on synthetic and real-world datasets demonstrate that MPHM achieves state-of-the-art performance while maintaining competitive model complexity. Future work will explore extending our multi-prior fusion paradigm to handle co-occurring degradations such as haze and illumination imbalance, further improving robustness under diverse weather conditions.

Acknowledgments

This research was financially supported by the National Natural Science Foundation of China (62501189, U23B2009), the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project (YQ2024F006) and Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-KF-24-09).

References

Bai, Y.; Wang, C.; Xie, S.; Dong, C.; Yuan, C.; and Wang, Z. 2025. Textir: A simple framework for text-based editable image restoration. *IEEE Transactions on Visualization and Computer Graphics*.

Chen, H.; Chen, X.; Lu, J.; and Li, Y. 2024. Rethinking multi-scale representations in deep deraining transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1046–1053.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12299–12310.

Chen, X.; Li, H.; Li, M.; and Pan, J. 2023a. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5896–5905.

Chen, X.; Pan, J.; and Dong, J. 2024. Bidirectional multi-scale implicit neural representations for image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25627–25636.

Chen, X.; Pan, J.; Dong, J.; and Tang, J. 2023b. Towards unified deep image deraining: A survey and a new benchmark. arXiv. *arXiv preprint arXiv:2310.03535*.

Chen, X.; Pan, J.; Lu, J.; Fan, Z.; and Li, H. 2023c. Hybrid cnn-transformer feature fusion for single image deraining. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 378–386.

Chen, Y.-L.; and Hsu, C.-T. 2013. A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In *Proceedings of the IEEE international conference on computer vision*, 1968–1975.

Dang, J.; Zheng, H.; Lai, J.; Yan, X.; and Guo, Y. 2023. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32: 3924–3938.

Dang, J.; Zheng, H.; Xu, X.; Wang, L.; Hu, Q.; and Guo, Y. 2024. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Fu, X.; Huang, J.; Zeng, D.; Huang, Y.; Ding, X.; and Paisley, J. 2017. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3855–3863.

Fu, X.; Qi, Q.; Zha, Z.-J.; Zhu, Y.; and Ding, X. 2021. Rain streak removal via dual graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1352–1360.

Gao, N.; Jiang, X.; Zhang, X.; and Deng, Y. 2024. Efficient Frequency-Domain Image Deraining with Contrastive Regularization. In *European Conference on Computer Vision*, 240–257. Springer.

Guo, Y.; Xiao, X.; Chang, Y.; Deng, S.; and Yan, L. 2023. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12097–12107.

Hong, S.; Yue, T.; You, Y.; Lv, Z.; Tang, X.; Hu, J.; and Yin, H. 2025. A Resilience Recovery Method for Complex Traffic Network Security Based on Trend Forecasting. *International Journal of Intelligent Systems*, 2025(1): 3715086.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Huang, B.; Luo, Y.; Ma, J.; and Jiang, J. 2020. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8346–8355.

Jin, X.; Shi, Y.; Xia, B.; and Yang, W. 2024. Llmra: Multi-modal large language model based restoration assistant. *arXiv preprint arXiv:2401.11401*.

Jocher, G. 2023. YOLOv8. Version 8.x.

Kang, L.-W.; Lin, C.-W.; and Fu, Y.-H. 2011. Automatic single-image-based rain streaks removal via image decomposition. *IEEE transactions on image processing*, 21(4): 1742–1755.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Li, D.; Liu, Y.; Fu, X.; Xu, S.; and Zha, Z.-J. 2024. Fourier-mamba: Fourier learning integration with state space models for image deraining. *arXiv preprint arXiv:2405.19450*.

Li, X.; Wu, J.; Lin, Z.; Liu, H.; and Zha, H. 2018. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, 254–269.

Li, Y.; Tan, R. T.; Guo, X.; Lu, J.; and Brown, M. S. 2016. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2736–2744.

Lin, X.; Yue, J.; Chan, K. C.; Qi, L.; Ren, C.; Pan, J.; and Yang, M.-H. 2023. Multi-task image restoration guided by robust dino features. *arXiv preprint arXiv:2312.01677*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I.; Hutter, F.; et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5: 5.

Luo, Y.; Xu, Y.; and Ji, H. 2015. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE international conference on computer vision*, 3397–3405.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Controlling vision-language models for multi-task image restoration. *arXiv preprint arXiv:2310.01018*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Paszke, A. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

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

Sun, S.; Ren, W.; Zhou, J.; Gan, J.; Wang, R.; and Cao, X. 2024. A hybrid transformer-mamba network for single image deraining. *arXiv preprint arXiv:2409.00410*.

Wang, C.; Xing, X.; Wu, Y.; Su, Z.; and Chen, J. 2020a. Dcsfn: Deep cross-scale fusion network for single image rain removal. In *Proceedings of the 28th ACM international conference on multimedia*, 1643–1651.

Wang, H.; Xie, Q.; Zhao, Q.; and Meng, D. 2020b. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3103–3112.

Wang, T.; Yang, X.; Xu, K.; Chen, S.; Zhang, Q.; and Lau, R. W. 2019. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12270–12279.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.

Xiao, J.; Fu, X.; Liu, A.; Wu, F.; and Zha, Z.-J. 2022. Image de-raining transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(11): 12978–12995.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1357–1366.

Yi, Q.; Li, J.; Dai, Q.; Fang, F.; Zhang, G.; and Zeng, T. 2021. Structure-preserving deraining with residue channel prior guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4238–4247.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.

Zhang, H.; and Patel, V. M. 2018. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 695–704.

Zhong, X.; Tu, S.; Ma, X.; Jiang, K.; Huang, W.; and Wang, Z. 2022. Rainy WCity: A Real Rainfall Dataset with Diverse Conditions for Semantic Driving Scene Understanding. In *IJCAI*, 1743–1749.