Rethinking Multi-Scale Representations in Deep Deraining Transformer

Hongming Chen[1], Xiang Chen[2*], Jiyang Lu[1], Yufeng Li[1*]

[1] College of Electronic Information Engineering, Shenyang Aerospace University
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology
{chenhongming,lujiyang1}@stu.sau.edu.cn, chenxiang@njust.edu.cn, liyufeng@sau.edu.cn

Abstract

Existing Transformer-based image deraining methods depend mostly on fixed single-input single-output U-Net architecture. In fact, this not only neglects the potentially explicit information from multiple image scales, but also lacks the capability of exploring the complementary implicit information across different scales. In this work, we rethink the multi-scale representations and design an effective multi-input multi-output framework that constructs intra- and inter-scale hierarchical modulation to better facilitate rain removal and help image restoration. We observe that rain levels reduce dramatically in coarser image scales, thus proposing to restore rain-free results from the coarsest scale to the finest scale in image pyramid inputs, which also alleviates the difficulty of model learning. Specifically, we integrate a sparsity-compensated Transformer block and a frequency-enhanced convolutional block into a coupled representation module, in order to jointly learn the intra-scale content-aware features. To facilitate representations learned at different scales to communicate with each other, we leverage a gated fusion module to adaptively aggregate the inter-scale spatial-aware features, which are rich in correlated information of rain appearances, leading to high-quality results. Extensive experiments demonstrate that our model achieves consistent gains on five benchmarks.

Introduction

Adverse visual conditions have become a major obstacle to the application of artificial intelligence, especially computer vision. As one of the harsh environments, rainy days have sparked a wave of research on low-level vision communities in recent years. The objective of single image deraining is to remove or reduce the undesired degradation caused by rain from input images, enhancing its visual quality and improving the accuracy of perception system (Chen et al. 2022).

Over the years, numerous techniques have been proposed to address the image deraining problem. Early prior-based methods (Kang, Lin, and Fu 2011; Chen and Hsu 2013; Luo, Xu, and Ji 2015; Li et al. 2016) utilize prior information about clean images or specific characteristics of rain streaks to guide the restoration process. Afterwards, the emergence of deep learning has attracted significant advancements in the field of image rain removal (Yang et al. 2020). CNN-based methods (Jiang et al. 2020; Chen et al. 2022) excel in their ability to learn intricate mappings between rainy and clean images, allowing them to effectively handle various shapes, sizes, and densities of rain streaks. More recently, Transformer-based methods (Chen et al. 2023a; Wang et al. 2022; Xiao et al. 2022; Chen et al. 2023b) have demonstrated remarkable performance in image deraining, primarily due to their ability to model non-local information, which is crucial for achieving high-quality image reconstruction.

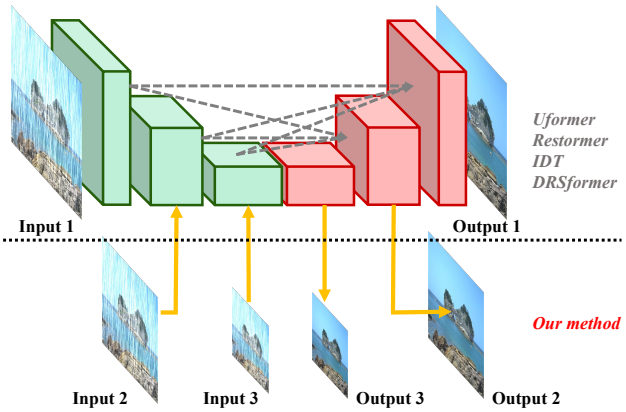


Figure 1: The architectures of deep deraining Transformers. Compared to existing models, our method first introduces a coarse-to-fine scheme to better capture multi-scale representations (intra-scale and inter-scale hierarchical modulation).

Despite achieving impressive performance, we note that existing deep image deraining Transformers depend mostly on single-input single-output architecture. Figure 1 provides a pipeline for network architecture. By this way, it not only neglects the potentially explicit information from multiple image scales, but also lacks the capability of exploring the complementary implicit information across different scales. In general, multi-scale visual information flow, encompassing feature representations at both small (global contextual) and large (local connectivity) scales, was frequently used in computer vision, particularly the solution of creating feature pyramids from image pyramid inputs (Chen, Zhu, and Gong 2017; Jiang et al. 2020; Mao et al. 2023). Thus, it is of great

*Corresponding author

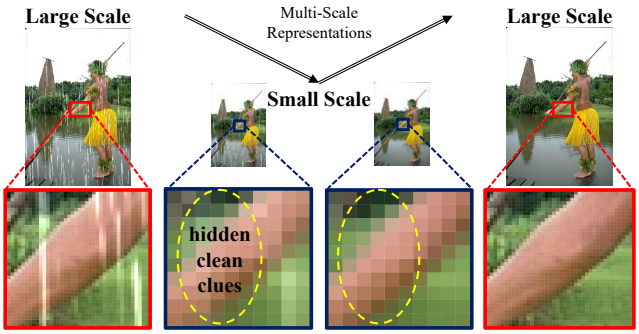


Figure 2: An example of rain image and clean image, and their coarser version at $1/4$ the scale. In these four images, we show the 10×10 patches at the same relative coordinates. Note that some signals similar to clean images are hidden on the some coarser scale of the rainy image.

interest to investigate multi-scale representations optimized for Transformer-based image deraining networks.

In fact, a single-scale representation tends to weaken hidden clean signals from other scales useful in feature learning. In Figure 2, we show multi-scale representations can be explored to "pull out" cleaner versions of the signal from the rainy image's coarser scales. We observe that rain levels reduce dramatically in coarser image scales. This permits the those potentially clean representations to "naturally emerge" in the rainy image at a coarser scale (Zontak, Mosseri, and Irani 2013; Michaeli and Irani 2014). With this finding, we are first attempt to formulate a new deep image Transformer architecture using a multi-input encoder and a multi-output decoder. Motivated by the coarse-to-fine scheme (Cho et al. 2021), we restore rain-free results from the coarsest scale to the finest scale in image pyramid inputs, which employs initial solutions estimated from those coarse scales to alleviate the difficulty of network learning. Here, we name our approach *Multi-Scale Deraining Transformer* (MSDT) design for not only excavating scale-specific discriminative feature, but also maximising scale-space complementary signals.

Specifically, our proposed MSDT is made up of 3 intra-scale branches that each learn one input image scale in the pyramid, as well as inter-scale branches that learn a complimentary combination of multi-scale rich representations. In the intra-scale branch, a sparsity-compensated Transformer block and a frequency-enhanced convolutional block is integrated into a coupled representation module. The former focuses on the most relevant global features via a neighbor softmax operator to facilitate a more accurate representation, while the latter integrates the local features information via a residual fourier transformation to better help image restoration. In addition, in the inter-scale branch, we leverage a gated fusion module to adaptively aggregate the multi-scale features generated by different encoder blocks to prompt the learning of various decoder blocks, thus boosting the final reconstruction performance. With this formulation, we allow all branches to be learned concurrently in an end-to-end fashion so as to achieve high-quality deraining outputs. Finally, extensive experiments show that our proposed MSDT delivers significant performance gains, exceeding the state-of-the-art Transformer-based approach DRSformer by 0.53 dB in PSNR on the real-world SPA-Data benchmark.

This paper makes the following contributions to the field:

- We rethink the multi-scale representations for single image deraining problem, and propose an effective end-to-end multi-input multi-output architecture to better facilitate rain removal in the richer scale space.
- We show that coupled representation modules can jointly learn the intra-scale content-aware features and gated fusion modules can be beneficial for the inter-scale spatial-aware features, in order to help hierarchical modulation.
- We perform comprehensive experiments to demonstrate the effectiveness of our method against the state-of-the-art Transformer-based image deraining approahces.

Related Work

We categorize existing methods into prior-based algorithms, CNN-based approaches and Transformer-based methods.

Prior-based methods. As image deraining is ill-posed, traditional approaches often employ hand-crafted priors based on image statistics to reconstruct images, *e.g.*, image decomposition (Kang, Lin, and Fu 2011), low-rank representation (Chen and Hsu 2013), discriminative sparse coding (Luo, Xu, and Ji 2015), and Gaussian mixture model (Li et al. 2016). However, these methods tend to rely on empirical observations and lead to complicated optimization problems.

CNN-based methods. Instead of manually designing image priors, CNN-based frameworks (Yang et al. 2020) have outperformed their conventional counterparts to achieve decent restoration performance. A wide array of network structures and designs have been effectively employed to significantly boost the capacity of end-to-end learning, *e.g.*, multi-scale (Jiang et al. 2020) and multi-stage fusion (Zamir et al. 2021). However, they have difficulty capturing non-local information due to the intrinsic limitations of convolution operators.

Transformer-based methods. Driven by the success of the Transformer network (Dosovitskiy et al. 2020), researchers have endeavored to replace CNN baselines with Transformers as the network backbone for vision tasks. In recent years, Transformer-based methods have been increasingly used for image restoration due to their superior capacity for modeling long-range dependencies. For example, Wang *et al*. (Wang et al. 2022) developed a general U-shaped Transformer architecture to solve image restoration problem. Zamir *et al*. (Zamir et al. 2022) designed an efficient Transformer model by estimating self-attention along the channel dimension, achieving remarkable performance. In the field of image rain removal, Xiao *et al*. (Xiao et al. 2022) first introduced the image deraining Transformer (IDT) using spatial-based and window-based self-attention modules. Recently, Chen *et al*. (Chen et al. 2023a) developed a sparse Transformer to make full use of the most useful features for better image restoration. However, existing Transformer-based methods are still limited to fixed single-input single-output U-Net baseline, lacking in exploring feature correlations in different image scale spaces. Our work will fill the gap in this research.

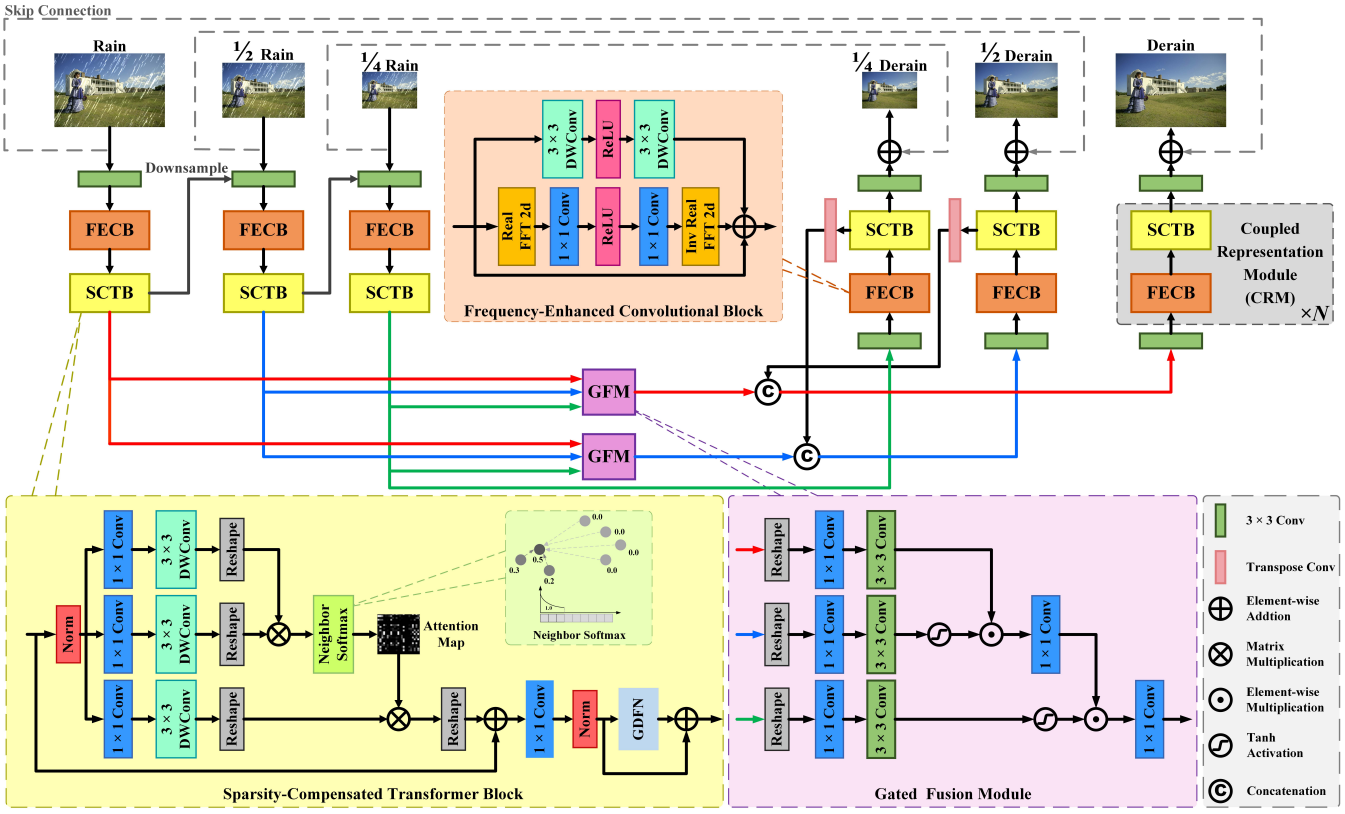


Figure 3: The architecture of the proposed multi-scale deraining Transformer (MSDT), which takes multi-input rainy images and generates multi-output derained images. It mainly contains (1) coupled representation module (CRM) with frequency-enhanced convolutional block (FECB) and sparsity-compensated Transformer block (SCTB), (2) gated fusion module (GFM).

Proposed Method

Our goal is to develop an effective deep image Transformer for high-quality image deraining, which fully makes full use of multi-scale features extracted from an input image. We describe each component of our network in the following.

Network Architecture

The overall framework of our proposed multi-scale deraining Transformer (MSDT) is presented in Figure 3, which is divided into 3 scales based on previous coarse-to-fine techniques (Kim, Lee, and Cho 2022). Given a input rainy image, our method first produces pyramid rainy images using the interpolation operator to down-sample the original rainy image into multiple scales, *i.e.*, 1/2 and 1/4. From the coarsest to the finest image scales, we designate each scale as $S3$, $S2$, and $S1$, respectively. The network takes pyramid rainy images as multi-inputs, and extracts the shallow features using a shallow 3×3 convolutional layers. Based on the initial features from each scale, N stacked coupled representation modules (CRMs) then perform the deep feature extraction and fusion of multi-scale rain information by inserting two parallel gated fusion modules (GFMs). On one side, intra-scale networks with the same structure are used to capture the most discriminative visual cues for each individual pyramid scale of rain appearances. On the other side, inter-scale networks are adopted for performing the discriminative feature selection and optimal integration of scale-specific representations from different scales. In the multi-input multi-output framework, the ultimate goal is to find relevant complementary combinations for feature selection at different scales, while optimizing the discriminative feature representation at each scale. In summary, compared to previous deep deraining Transformers, our design can more effectively find more potential clean signals in multi-scale feature learning.

Model Optimization

In order to supervise the learning process of the network, we choose three kinds of loss functions as training objectives to guide the optimization, which are calculated as follows:

$$\mathcal{L}_{msc} = \sum_{k=1}^{3} \sqrt{\left\| \hat{\mathbf{B}}_k - \mathbf{B}_k \right\|^2 + \varepsilon^2}, \tag{1}$$

where $\hat{\mathbf{B}}_k$ and $\mathbf{B}_k$ denote the k-th scale reconstructed image and ground-truth image, respectively. $\mathcal{L}_{msc}$ represents the Multi-Scale Charbonnier (MSC) loss (Charbonnier et al. 1994). Here, the penalty coefficient ε is set to 10^{-3}.

$$\mathcal{L}_{msed} = \sum_{k=1}^{3} \sqrt{\left\| \Delta\left(\hat{\mathbf{B}}_k\right) - \Delta\left(\mathbf{B}_k\right) \right\|^2 + \varepsilon^2}, \tag{2}$$

where $\mathcal{L}_{msed}$ is the Multi-Scale Edge (MSED) loss (Zamir et al. 2021). Here, Δ represents the Laplacian operator.

$$\mathcal{L}_{msfr} = \sum_{k=1}^{3} \left\| \mathcal{FT}\left(\hat{\mathbf{B}}_k\right) - \mathcal{FT}(\mathbf{B}) \right\|_1, \tag{3}$$

where $\mathcal{L}_{msfr}$ is the Multi-Scale Frequency Reconstruction (MSFR) loss (Cho et al. 2021). $\mathcal{FT}$ represents the Fourier transform operator to obtain the frequency domain of the original image. Finally, the overall loss $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{msc} + \lambda_2 \mathcal{L}_{msed} + \lambda_3 \mathcal{L}_{msfr}, \tag{4}$$

where the trade-off weight λ_1, λ_2 and λ_3 are empirically set to 1, 0.05, and 0.01 as in (Mao et al. 2023), respectively.

Coupled Representation Module (CRM)

In order to better satisfy rain removal, both local and global information representations are increasingly indispensable (Jiang et al. 2022; Chen et al. 2023b). Hence, we propose a coupled representation module (CRM) that combines unique advantages of CNN and Transformer. Here, our developed CRM consists two main designs: frequency-enhanced convolutional block (FECB) and sparsity-compensated Transformer block (SCTB). In this way, integrating sparsity and frequency guidance into multi-scale architecture is of great importance for boosting image restoration performance. We will describe the details about these two components.

Frequency-enhanced convolutional block (FECB) is introduced to improve the locality of the network in the frequency domain (Mao et al. 2023). To recover the texture details of the background images, we first use Fast Fourier Transform (FFT) (Nussbaumer and Nussbaumer 1981) to extract high-frequency components. Due to the symmetric nature of the FFT, we only use the Real FFT in order to reduce the computational overhead. Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ be the input feature of FECB. We first transform $\mathbf{X}$ into a frequency domain with $2d$ Real FFT to obtain $\mathcal{F}(\mathbf{X}) \in \mathbb{R}^{H \times \frac{W}{2} \times C}$. Then, the real and imaginary parts of $\mathcal{F}(\mathbf{X})$ are concatenated along the channel dimension to obtain $\mathbf{Y} \in \mathbb{R}^{H \times \frac{W}{2} \times 2C}$. Next, $\mathbf{Y}$ is applied to two 1×1 convolutions and one nonlinear function. Finally, $\mathbf{Y}$ is recovered to the spatial structure with the inverse $2d$ Real FFT, we define it as $\mathbf{Y}_{fft} = \mathcal{F}^{-1}(\mathbf{Y}) \in \mathbb{R}^{H \times W \times C}$. Similar to (Mao et al. 2023), we also add residual paths to boost inter-scale feature propagation.

Sparsity-compensated Transformer block (SCTB) is introduced to focus on the most relevant non-local information in each scale, which enables much accurate representations. In each SCTB, given the input features at the (l-1)-th block $\mathbf{X}_{l-1}$, the encoding procedures of SCTB can be formulated as:

$$\mathbf{X}'_l = \mathbf{X}_{l-1} + \text{NSSA}\left(\text{LN}\left(\mathbf{X}_{l-1}\right)\right), \tag{5}$$

$$\mathbf{X}_l = \mathbf{X}'_l + \text{GDFN}\left(\text{LN}\left(\mathbf{X}'_l\right)\right), \tag{6}$$

where $\mathbf{X}'_l$ and $\mathbf{X}_l$ represent the outputs from the neighbor softmax self-attention (NSSA) and gated-Dconv feed-forward network (GDFN) (Zamir et al. 2022). Here, LN refers to the layer normalization.

Motivated by (Chen et al. 2023a), we note that the softmax normalization in the standard Transformer is performed on all the input tokens, thus redundant irrelevant representations will interfere with the feature aggregation and distract the attention. To alleviate this problem, we develop the neighbor softmax to replace the normal softmax. Specifically, we first encode channel-wise context by applying 1×1 convolutions followed by 3×3 depth-wise convolutions. Given the query Q, key K, and value V, we generate the attention values P of all pixel pairs between Q and K:

$$P = \frac{QK^{\mathrm{T}}}{\sqrt{d}}, \tag{7}$$

where $d = C/k$ is the head dimension and k is the head number. We presum that two tokens are likely to be relevant if they are latent neighbours to one another in feature space. Here, we propose a simple but effective masking function $\mathcal{M}(\cdot)$ is performed upon P to select the top-k neighbors from each row of the similarity matrix. For other elements that are smaller than threshold, we replace them with 0. By adding this mask $\mathcal{M}(\cdot)$ to the regular softmax function, we achieve sparse self-attention only occurring in neighbors, which is calculated by

$$\mathcal{M}(P,k)_{ij} = \begin{cases} P_{ij} & \text{if } P_{ij} \geq \text{ threshold} \\ 0 & \text{if } P_{ij} < \text{ threshold} \end{cases}, \tag{8}$$

where threshold is a k^{th} largest value of row. Considering the discrepancies in the degree of rain degradation at different image scales, we adopt different thresholds at each scale branches to better maximise scale-specific feature. Due to the partial attention value being set to zero, the relationship is constrained to the relevant neighbors, making the aggregation of features within the scale more focused and robust.

Gated Fusion Module (GFM)

For feature fusion, scale-specific representations from different scales that have different spatial resolutions are typically scaled by down-sampling or up-sampling procedures to the same spatial resolution. The final image restoration may be impacted by these resizing operations because they could result in the loss of some crucial structural elements. To this end, we formulate a gated fusion module to adaptively aggregate the inter-scale information flow. Given input features at three scales, the process of GFM can be expressed as:

$$F_{GFM} = F_{1\times1}\left(\text{Gate}\left(F_{1\times1}(\text{Gate}(S3, S2)), S1\right)\right), \tag{9}$$

where $F_{1\times1}$ denotes a 1×1 convolutional layer. Here, we adopt dual gate units to dynamically control hierarchical information. Each gate unit $Gate(\cdot)$ is formulated as:

$$\text{Gate}(\mathbf{X}, \mathbf{Y}) = \sigma\left(F_{3\times3}\left(F_{1\times1}(\mathbf{Y})\right)\right) \odot F_{3\times3}\left(F_{1\times1}(\mathbf{X})\right), \tag{10}$$

where σ refers to the tanh activation function, and $\odot$ represents the element-wise multiplication. The output of the GFM is delivered to its corresponding decoder part. By incorporating the GFM, our proposed model effectively ensure synergistically correlated multi-scale feature learning. In what follows, we will show the effectiveness of these design choices in the experimental section.

Datasets		Rain200L		Rain200H		DID-Data		DDN-Data		SPA-Data	
Metrics		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Prior-based methods	DSC	27.16	0.8663	14.73	0.3815	24.24	0.8279	27.31	0.8373	34.95	0.9416
	GMM	28.66	0.8652	14.50	0.4164	25.81	0.8344	27.55	0.8479	34.30	0.9428
CNN-based methods	DDN	34.68	0.9671	26.05	0.8056	30.97	0.9116	30.00	0.9041	36.16	0.9457
	RESCAN	36.09	0.9697	26.75	0.8353	33.38	0.9417	31.94	0.9345	38.11	0.9707
	PReNet	37.80	0.9814	29.04	0.8991	33.17	0.9481	32.60	0.9459	40.16	0.9816
	MSPFN	38.58	0.9827	29.36	0.9034	33.72	0.9550	32.99	0.9333	43.43	0.9843
	RCDNet	39.17	0.9885	30.24	0.9048	34.08	0.9532	33.04	0.9472	43.36	0.9831
	MPRNet	39.47	0.9825	30.67	0.9110	33.99	0.9590	33.10	0.9347	43.64	0.9844
	DualGCN	40.73	0.9886	31.15	0.9125	34.37	0.9620	33.01	0.9489	44.18	0.9902
	SPDNet	40.50	0.9875	31.28	0.9207	34.57	0.9560	33.15	0.9457	43.20	0.9871
Transformer-based methods	Uformer	40.20	0.9860	30.80	0.9105	35.02	0.9621	33.95	0.9545	46.13	0.9913
	Restormer	40.99	0.9890	32.00	0.9329	35.29	0.9641	34.20	0.9571	47.98	0.9921
	IDT	40.74	0.9884	32.10	0.9344	34.89	0.9623	33.84	0.9549	47.35	**0.9930**
	DRSformer	41.23	0.9894	32.17	0.9326	35.35	0.9646	34.35	0.9588	48.54	0.9924
	Ours	**41.75**	**0.9904**	**32.45**	**0.9379**	**35.37**	**0.9652**	**34.36**	**0.9593**	**49.07**	0.9926

Table 1: Comparison of quantitative results on five benchmarks. Bold and underline indicate the best and second-best results.

Experiments

In this section, comprehensive image deraining experiments is performed on commonly used benchmark datasets to evaluate the effectiveness of the proposed method. The training code and test model will be available to the public.

Experimental Setup

Datasets and metrics. We evaluate the performance of our model on five publicly rain streak datasets: Rain200L (Yang et al. 2017), Rain200H (Yang et al. 2017), DID-Data (Zhang and Patel 2018), DDN-Data (Fu et al. 2017), and SPA-Data (Wang et al. 2019). Rain200L and Rain200H comprise 1,800 synthetic rainy images for training, along with 200 images designated for testing. DID-Data and DDN-Data comprise 12,000 and 12,600 synthetic images, featuring distinct rain directions and density levels. Each dataset includes 1,200 and 1,400 rainy images specifically designated for testing. In addition, SPA-Data is a large-scale real-world rain benchmark, encompassing 638,492 image pairs for training, alongside 1,000 image pairs designated for testing. Given the availability of ground truths, we employ two widely used metrics for quantitative comparison: Peak Signal to Noise Ratio (PSNR) (Huynh-Thu and Ghanbari 2008) and Structural Similarity (SSIM) (Wang et al. 2004). Following previous deraining methods (Fu et al. 2023; Chen et al. 2023a), we calculate those metrics in Y channel of YCbCr space.

Comparison methods. We compare our proposed approach with diverse state-of-the-art image deraining baselines, including two prior-based algorithms (*i.e.*, DSC (Luo, Xu, and Ji 2015) and GMM (Li et al. 2016)), eight CNN-based approaches (*i.e.*, DDN (Fu et al. 2017), RESCAN (Li et al. 2018), PReNet (Ren et al. 2019), MSPFN (Jiang et al. 2020), RCDNet (Wang et al. 2020), MPRNet (Zamir et al. 2021), DualGCN (Fu et al. 2021), and SPDNet (Yi et al. 2021)) and four Transformer-based networks (*i.e.*, Uformer (Wang et al. 2022), Restormer (Zamir et al. 2022), IDT (Xiao et al. 2022) and DRSformer (Chen et al. 2023a)). To ensure fair comparison, we refer to their online experimental data provided by (Chen et al. 2023a) using same evaluation protocol.

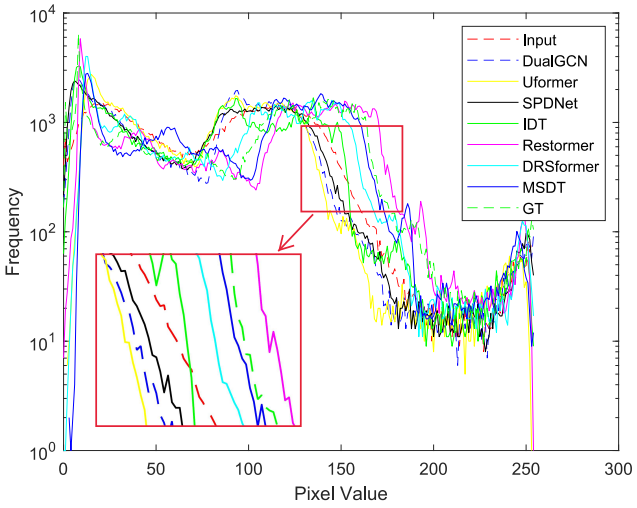
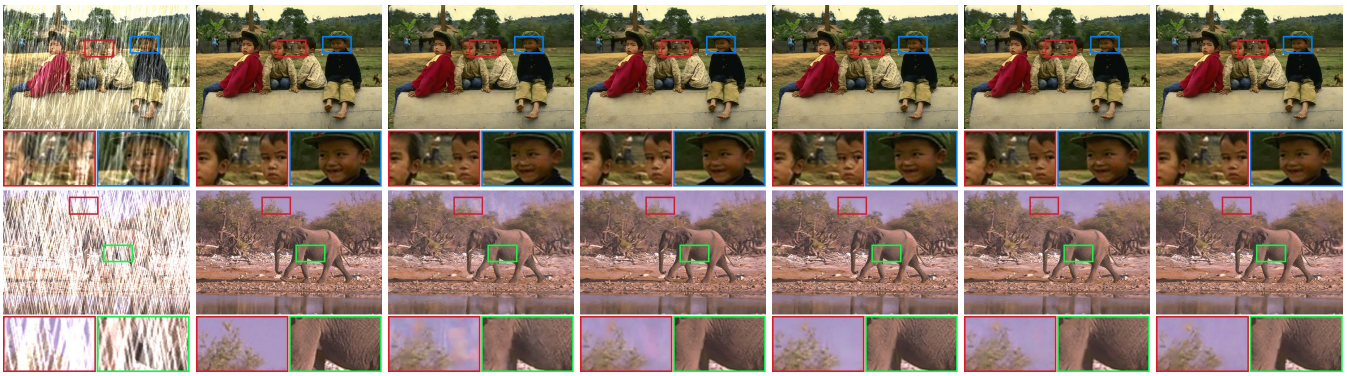


Figure 4: The average fitting results of the synthetic datasets based on the histogram curve of Y channel in YCbCr space, suggesting that our proposed MSDT leads in greater fitting accuracy to the ground-truths distribution than other comparison approaches.

Implementation details. During training, the proposed network is implemented in PyTorch framework using Adam optimizer with a learning rate of 2×10^{-4} to minimize $\mathcal{L}_{total}$. The final learning rate is steadily decreased to 1×10^{-4} using the cosine annealing strategy (Loshchilov and Hutter 2016). For Rain200L, Rain200H, DID-Data and DDN-Data, 500 epochs are trained, while SPA-Data is trained for 5 epochs. For data augmentation, we also randomly adopt horizontal and vertical flips. In our model, we adopt a stack of 8 CRMs (*i.e.*, $N = 8$ in Figure 3. We set the thresholds of SCTB in $S1$, $S2$, and $S3$ to 0.6, 0.7, and 0.8, respectively. The setting of GDFN in SCTB is consistent with (Zamir et al. 2022). We run all of our experiments with batch size of 2 and patch size of 256 on one NVIDIA GeForce RTX 4090 GPU (24G). For testing, sliding window slicing crop method is employed.



(a) Rainy Input (b) GT (c) SPDNet (d) Restormer (e) IDT (f) DRSformer (g) **Ours**

Figure 5: Visual comparison on the Rain200H dataset. Best viewed by zooming in the figures on high-resolution displays.

	Uformer	Restormer	IDT	DRSformer	Ours
#P	50.8	26.1	16.4	33.7	16.6
#F	45.9	174.7	61.9	242.9	129.9
#R	0.19	0.28	0.28	0.31	0.09

Table 2: Comparison of model efficiency on a 256×256 image. "#P", "#F" and "#R" represent the number of trainable parameters (in M), FLOPs (in G) and inference time (in second), respectively.

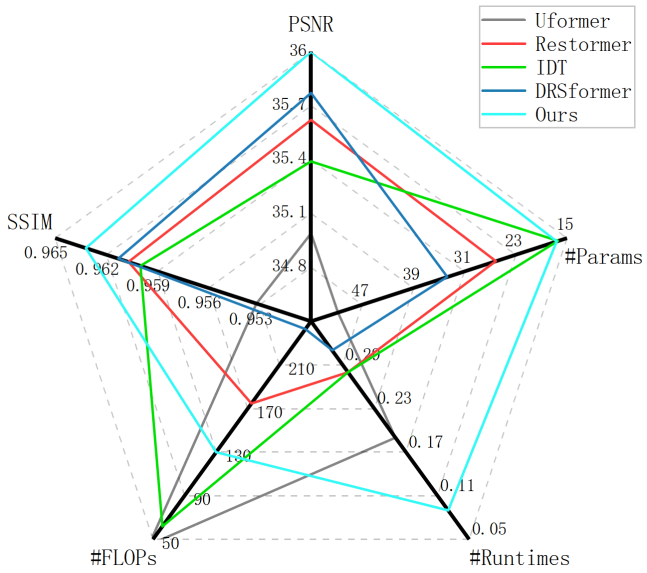


Figure 6: Five dimensional radar map of comprehensive capability of deep deraining Transformers, including PSNR, SSIM, #Params, #FLOPs and #Runtimes.

Experimental Results

Synthetic datasets. Table 1 compares our method with 14 representative and state-of-the-art deraining approaches. We can clearly see that our proposed MSDT consistently outperforms all the other baselines in terms of PSNR and SSIM values thanks to our multi-scale architecture, especially exceeding the recent Transformer-based approach DRSformer (Chen et al. 2023a) by 0.52 dB in PSNR on the Rain200L benchmark. Following (Jiang et al. 2022), we further show the average fitting results of the synthetic datasets based on the histogram curve of Y channel in the YCbCr space, indicating that our deraining results of MSDT are close to the distribution of ground-truths images (Figure 4). The high-quality performance gains also confirm that our framework opens up a new perspective for deep image deraining architecture. Furthermore, Figure 5 shows a qualitative comparison on the Rain200H dataset. As can be seen, our method yields cleaner results compared to other approaches, which is also consistent with the quantitative values.

Real-world datasets. To facilitate real-world performance analysis, we conduct experiments on the SPA-Data. The last column of Table 1 records the corresponding the quantitative results. As expected, it can be seen that the PSNR and SSIM values of our method still maintain a high level of competitiveness. Meanwhile, we present several examples of visual comparison results in Figure 7. According to the ground truth, all of the other efforts undertake undesirable results in rain removal or detail restoration, and our proposed MSDT outperforms the others. This further validates the performance advantage of our multi-scale coarse-to-fine feature learning method over single-scale feature learning based single-input single-output U-Net framework.

Model efficiency. In Table 2, we also compare the computational complexity of different deep deraining Transformer-based methods, including the number of trainable parameters, FLOPs and inference time on a 256×256 image. To intuitively compare the comprehensive capabilities of recent state-of-the-art methods, we present a radar map of model capability in Figure 6. Compared to other deep Transformer-based methods, our model achieves competitive results in four dimensions except for #FLOPs, which implies the potential of our method to become a pentagon warrior.

Ablation Studies

In what follows, we adopt the challenging Rain200H dataset to conduct ablation studies for analysis and discussions.

(a) Rainy Input (b) GT (c) DualGCN (d) SPDNet (e) Restormer (f) IDT (g) DRSformer (h) **Ours**

Figure 7: Visual comparison on the SPA-Data dataset. Best viewed by zooming in the figures on high-resolution displays.

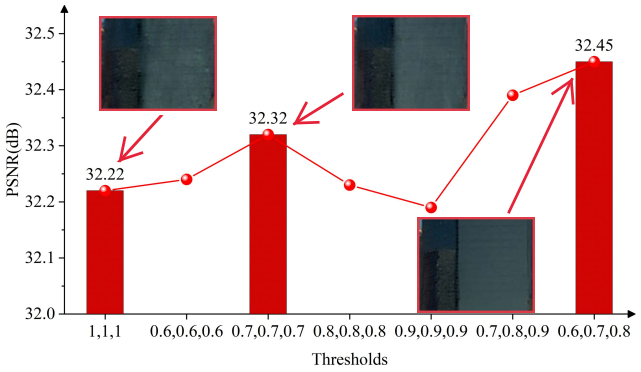


Figure 8: Ablation analysis for different thresholds (from left to right: S1, S2, S3) in the SCTB.

Scale Level	S1	S1+S2	S1+S2+S3
PSNR	32.08	32.20	**32.45**

Table 3: Ablation analysis for different levels of image scale.

Effectiveness of multi-scale. To investigate the effect about different levels of image scale, Table 3 reports the PSNR/SSIM values of corresponding models. Compared to performing single-scale deraining, richer multi-scale representations can bring a great contribution to the baseline model. Because of the negligible rain effect at a coarse scale, it can predict a more accurate result, which acts as an initial solution for a finer scale, resulting in a higher-quality deraining result.

Effectiveness of CRM. The influence of threshold, the crucial hyper-parameter for our proposed SCTB, is investigated in Figure 8. As shown, our developed neighbor softmax outperforms normal softmax function (*i.e.*, set all thresholds to 1) in terms of PSNR, which suggests that paying attention to relevant neighbors only leads to better feature representations. In addition, setting different thresholds for different scales can be more conducive to global information aggregation. We further demonstrate the effectiveness of FECB in Table 4. Through the quantitative results, the recovered results of the model with FECB in the CRM tend to be better.

Method	w/o GFM	w/o FECB	Ours
PSNR	32.33	31.55	**32.45**

Table 4: Ablation analysis for GFM and FECB in the CRM.

Loss	w/ MSC	w/o MSFR	w/o MSED	Ours
PSNR	32.17	32.32	32.35	**32.45**

Table 5: Ablation analysis for different loss functions.

Effectiveness of GFM. We further analyze the effectiveness of GFM in Table 4. Compared to direct feature concatenation (*i.e.*, without GFM), our designed solution tends to be more effective for combining multi-scale representations.

Effectiveness of loss functions. The quantitative comparisons in Table 5 show the significance of using hybrid losses, which indicates that each component we considered has its own contribution to the final deraining performance.

Concluding Remarks

This paper first proposes a high-quality multi-scale deraining Transformer (MSDT) for further boosting image restoration performance. In our designed architecture, we show that coupled representation modules can jointly learn the intra-scale content-aware features and gated fusion modules can be beneficial for the inter scale spatial-aware features. Experiments demonstrate the superiority of our method over the state-of-the-arts. We hope that our proposed multi-input multi-output architecture can provide a new perspective and solution for exploring more related low-level vision tasks.

Acknowledgements

This work has been supported in part by the Liaoning Provincial Applied Basic Research Project under Grant 2022JH2/101300247, and Shenyang Science and Technology Project under Grant 23-503-6-18.

References

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, volume 2, 168–172.

Chen, X.; Li, H.; Li, M.; and Pan, J. 2023a. Learning A Sparse Transformer Network for Effective Image Deraining. In *CVPR*, 5896–5905.

Chen, X.; Pan, J.; Jiang, K.; Li, Y.; Huang, Y.; Kong, C.; Dai, L.; and Fan, Z. 2022. Unpaired deep image deraining using dual contrastive learning. In *CVPR*, 2017–2026.

Chen, X.; Pan, J.; Lu, J.; Fan, Z.; and Li, H. 2023b. Hybrid cnn-transformer feature fusion for single image deraining. In *AAAI*, volume 37, 378–386.

Chen, Y.; Zhu, X.; and Gong, S. 2017. Person re-identification by deep learning multi-scale representations. In *ICCVW*, 2590–2600.

Chen, Y.-L.; and Hsu, C.-T. 2013. A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In *ICCV*, 1968–1975.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 4641–4650.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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 *CVPR*, 3855–3863.

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

Fu, X.; Xiao, J.; Zhu, Y.; Liu, A.; Wu, F.; and Zha, Z.-J. 2023. Continual image deraining with hypergraph convolutional networks. *IEEE TPAMI*.

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.; Chen, C.; Wang, Z.; Cui, L.; and Lin, C.-W. 2022. Magic ELF: Image deraining meets association learning and transformer. *ACM MM*.

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 *CVPR*, 8346–8355.

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

Kim, K.; Lee, S.; and Cho, S. 2022. Mssnet: Multi-scale-stage network for single image deblurring. In *ECCV*, 524–539.

Li, X.; Wu, J.; Lin, Z.; Liu, H.; and Zha, H. 2018. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 254–269.

Li, Y.; Tan, R. T.; Guo, X.; Lu, J.; and Brown, M. S. 2016. Rain streak removal using layer priors. In *CVPR*, 2736–2744.

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

Luo, Y.; Xu, Y.; and Ji, H. 2015. Removing rain from a single image via discriminative sparse coding. In *ICCV*, 3397–3405.

Mao, X.; Liu, Y.; Liu, F.; Li, Q.; Shen, W.; and Wang, Y. 2023. Intriguing findings of frequency selection for image deblurring. In *AAAI*, volume 37, 1905–1913.

Michaeli, T.; and Irani, M. 2014. Blind deblurring using internal patch recurrence. In *ECCV*, 783–798.

Nussbaumer, H. J.; and Nussbaumer, H. J. 1981. *The fast Fourier transform*.

Ren, D.; Zuo, W.; Hu, Q.; Zhu, P.; and Meng, D. 2019. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 3937–3946.

Wang, H.; Xie, Q.; Zhao, Q.; and Meng, D. 2020. A model-driven deep neural network for single image rain removal. In *CVPR*, 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 *CVPR*, 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 TIP*, 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 *CVPR*, 17683–17693.

Xiao, J.; Fu, X.; Liu, A.; Wu, F.; and Zha, Z.-J. 2022. Image de-raining transformer. *IEEE TPAMI*.

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 *CVPR*, 1357–1366.

Yang, W.; Tan, R. T.; Wang, S.; Fang, Y.; and Liu, J. 2020. Single image deraining: From model-based to data-driven and beyond. *IEEE TPAMI*, 43(11): 4059–4077.

Yi, Q.; Li, J.; Dai, Q.; Fang, F.; Zhang, G.; and Zeng, T. 2021. Structure-preserving deraining with residue channel prior guidance. In *ICCV*, 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 *CVPR*, 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 *CVPR*, 14821–14831.

Zhang, H.; and Patel, V. M. 2018. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, 695–704.

Zontak, M.; Mosseri, I.; and Irani, M. 2013. Separating signal from noise using patch recurrence across scales. In *CVPR*, 1195–1202.