

Article

Image Deraining Using Transformer Network with Sparse Non-Local Self-Attention

Xueying Zhao and Yufeng Li *

School of Electronic Information Engineering, Shenyang Aerospace University, Shenyang 110136, China; 20012324@sau.edu.cn

* Correspondence: liyufeng@sau.edu.cn

Abstract

In recent years, Transformer architectures have excelled at modeling non-local information. This makes them suitable for image deraining. However, existing methods use dense self-attention. They compute all similarities between query and key tokens. This is inefficient. In practice, this approach can lead to the neglect of the most relevant information and result in a blurring effect of irrelevant representations during the feature aggregation process. To address this issue, this paper proposes an image deraining Transformer based on sparse non-local self-attention. The core of the network consists of multiple non-local feature extraction modules, primarily comprising a sparse self-attention network and a sparse feedforward network along the channel dimension. Specifically, we implement sparse attention by selecting the most useful similarities based on Top-k approximations. Furthermore, we have developed a sparse feedforward network to achieve more accurate representations for high-quality preservation results. Extensive experiments on benchmark datasets have demonstrated the effectiveness of our proposed method.

Keywords: transformer; image deraining; self-attention; sparse attention

1. Introduction

Single-image deraining has risen as a classic signal processing task in the past decade. It aims to recover clear backgrounds from rainy scenes [1–6]. Compared to prior-based algorithms [1,2], deep learning has achieved significant success in this domain in recent years, with numerous deraining methods proposed [7–12]. Convolutional Neural Networks (CNNs) have strong learning capabilities. Researchers use them to model correlations between rainy and clear images. These CNN-based networks have demonstrated commendable performance in image restoration.

However, due to the limited receptive field and the interaction of convolutions, CNNs have a constrained ability to model long-range dependency information [13,14]. To address this bottleneck, recent studies have begun to incorporate Transformers into various image restoration tasks [15–17].

Recently, Vision Transformers have been applied to image deraining and have achieved significant progress. This is due to their ability to model non-local information, which is crucial for better image reconstruction. Zhou et al. [16] proposed a dual-path multi-scale Transformer for image deraining. This model combines coarse-to-fine and multi-patch stacked strategies, which enable it to exploit multi-scale information for high-quality image reconstruction. Wu et al. [17] employed a dual-branch architecture and soft prompt learning for image deraining. This method dynamically encodes rain degradation information and



Academic Editors: Selene Tomassini and M. Ali Akber Dewan

Received: 21 January 2026

Revised: 15 February 2026

Accepted: 16 February 2026

Published: 20 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

adopts multi-scale feature prompt fusion. It thus leverages the multi-scale properties of rain streaks to achieve high-quality image reconstruction. Although these methods have improved deraining performance, they still fall short when it comes to restoring image details. In the computation of similarity within Transformers, the standard softmax normalization operation retains all similarities between Query and Key features. However, not all features between Queries and Keys are closely related. Thus, using all similarities in non-local self-attention can introduce some irrelevant features, leading to a loss of detail in the final image restoration.

These methods achieve initial success. However, they struggle with fine details. Heavy rain makes this worse. Dense rain streaks cause severe degradation. These methods overlook a key problem. Dense self-attention computation within the Transformer affects recovery quality. It is known that in standard self-attention [18], assigning attention to all query-key pairs is the default operation. The most recent works [15–17] have attempted to apply this dense self-attention directly to Transformer-based image restorers. Naturally, this results in an inability to focus on the most relevant parts during feature aggregation. In other words, the integration of irrelevant features may have a blurring effect on the reconstruction results. Therefore, this prompts us to explore more effective ways of self-attention within the structure of the Transformer.

To this end, we propose an effective Image Deraining Transformer with channel-wise sparse self-attention (CSSA) and sparse feedforward network (SFFN). This model integrates features from both channel and spatial dimensions and adopts a sparse strategy to more efficiently facilitate the reconstruction of clear images. Unlike previous dense self-attention approaches, we also implement sparse self-attention by selecting the most useful similarities based on Top-k approximations. Furthermore, we have developed an efficient sparse feedforward network to simultaneously achieve spatial and channel integration within the Transformer. With the above design, our method promotes more effective feature aggregation and generates more accurate representations, thereby enhancing rain removal, as shown in Figure 1.

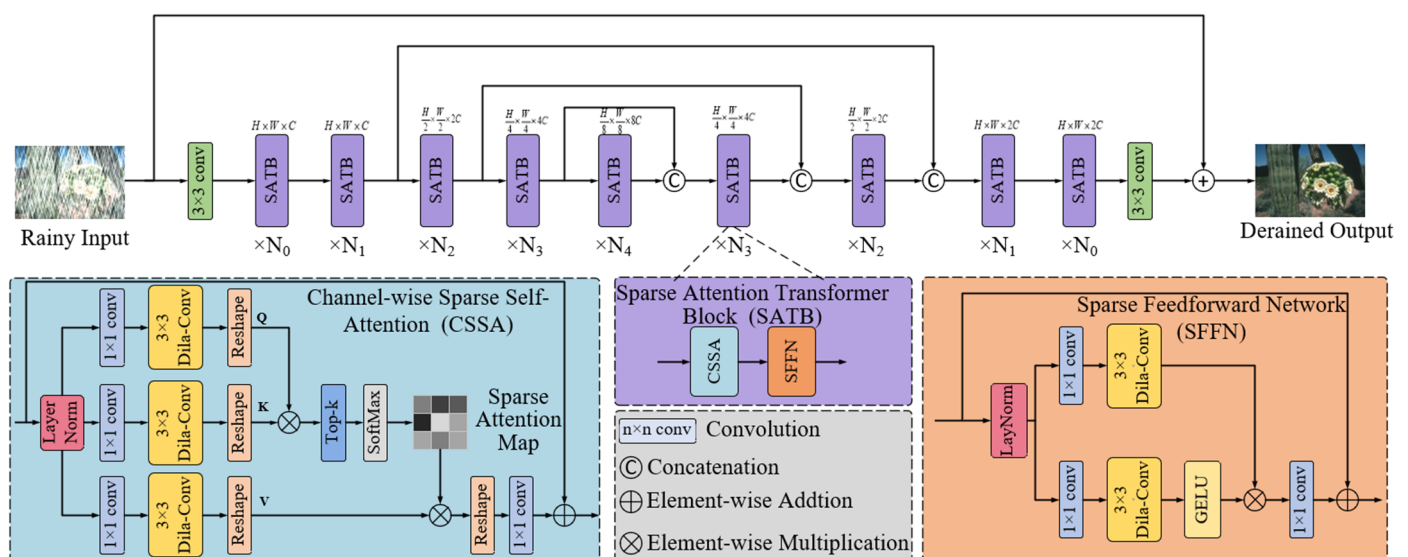


Figure 1. The overall of the proposed image deraining Transformer with sparse attention for single image rain removal.

The main contributions of this paper are as follows:

- (1) We have designed an effective sparse self-attention algorithm within the Transformer that can efficiently explore both local and global features, thereby enabling high-quality image reconstruction.
- (2) We have designed a sparse feedforward network within the Transformer that facilitates the integration of spatial and channel-wise rain information during feature transformation.
- (3) Extensive experimental results on common benchmarks demonstrate that our method is also highly competitive compared to the state-of-the-art deraining algorithms.

2. Related Work

2.1. Non-Local Self-Attention

Non-local self-attention is a core component of the Vision Transformer. It outperforms traditional local self-attention with remarkable improvements. Local self-attention restricts the computation of attention weights to adjacent positions. By contrast, non-local self-attention allows the model to capture information across the entire global scope. This capability is crucial for capturing long-range dependencies within an image, thereby enhancing the model's ability to perceive and integrate global contextual information [19]. In traditional self-attention, the limited scope of attention to nearby positions constrains the model's understanding of the overall image context. Non-local self-attention addresses this limitation by introducing an additional attention matrix. This design enables the model to compute attention between all positions in the image.

In recent years, Vision Transformer architectures integrated with non-local self-attention have been widely applied to various image restoration tasks. These tasks include image denoising [20], image super-resolution [21], and image deblurring [22]. This approach to non-local feature modeling offers several key advantages. First, it grants access to global contextual information. It allows the model to capture dependencies over a broader scope, thus better understanding the intricate details in an image. Second, it effectively handles long-range dependencies. It mitigates the problem of limited information propagation, which is especially critical for complex and variable image reconstruction tasks. Third, the non-local self-attention mechanism has strong adaptivity. It can automatically adjust attention weights according to the content of the input image. This adaptability allows the model to tailor its processing to different restoration tasks and data characteristics, thereby improving overall performance and robustness.

Non-local self-attention leverages these advantages to boost the capacity of Vision Transformers for high-quality image restoration. This makes them highly effective tools in the field of image processing. The integration of non-local self-attention mechanisms not only mitigates the limitations of traditional local self-attention. It also paves new ways for the advancement of various image restoration applications.

2.2. Sparse Representation

Sparse representation is a powerful method used in neural networks to reduce computational and storage costs. At the same time, it helps improve model generalization and robustness. This technique selectively activates only the neurons relevant to the current task. In this way, it enables more efficient learning of key features and reduces interference from irrelevant information. The application of sparse representation spans multiple domains and tasks, demonstrating its versatility and effectiveness. In the context of image super-resolution, Hu et al. [23] propose a global sparse attention network. This network groups similar features through spherical locality sensitive hashing and computes sparse attention within groups. It reduces computational complexity from quadratic to asymptotically linear with respect to spatial size, while capturing effective global information. As a result,

it achieves efficient super-resolution reconstruction for remote sensing images. Similarly, Garas Gendy [24] proposed a degradation-aware Swin Transformer with sparse attention for blind super-resolution. This method integrates Swin Transformer local attention, a non-local sparse attention global mechanism, and degradation-aware convolution. It can effectively handle multi-modal, spatially variant, and unknown distribution degradations in real-world scenarios.

Unlike these prior methods, this paper focuses on the image deraining task and employs a combination of Top-k approximation calculations and dilated convolutions to achieve sparsity. The Top-k approximation selectively retains the most significant similarities, thereby focusing computational resources on the most relevant features. Dilated convolutions can expand the receptive field without increasing computational load. They enable the network to capture more contextual information with fewer parameters. This combination improves both model efficiency and image restoration quality. It can effectively focus on key details and filter out irrelevant information.

3. Proposed Method

3.1. Overall Framework

This paper introduces an innovative image deraining algorithm that leverages sparse non-local self-attention mechanisms, as depicted in Figure 1. The network is designed as a symmetric layered encoder–decoder structure with long skip connections, crucial for preserving spatial information throughout the network. The backbone comprises sequentially stacked Transformer-based non-local feature extraction modules, each operating at specific spatial resolutions. During the encoding phase, spatial resolution is progressively halved at each layer while the channel dimension is expanded, allowing the network to capture increasingly complex features. Conversely, in the decoding phase, spatial resolution is restored, and the channel dimension is reduced, refining the feature maps to reconstruct the final image. For effective downsampling and upsampling, we employ pixel shuffling operations as suggested by prior research [15]. Pixel shuffling aids in efficiently managing spatial and channel dimensions, ensuring accurate transformations. The final output is restored to its original size through a 3×3 convolutional layer, ensuring precise reconstruction. This detailed and methodical design allows our network to effectively address the complexities of single-image deraining, achieving superior performance in removing rain streaks and restoring clear images. Extensive experiments on various benchmarks have demonstrated the effectiveness of our approach, highlighting its ability to produce high-quality, rain-free images. The combination of sparse self-attention and Transformer-based architecture makes our method robust and efficient, paving the way for further advancements in image restoration tasks.

The introduction of sparse attention with Top-k is aimed at focusing on the most relevant non-local information at each scale, thereby enabling more accurate representations. Within each Transformer module, given the input features at the $(l-1)$ -th block X_{l-1} , the encoding procedures of the blocks can be formulated as:

$$X'_l = X_{l-1} + \text{CSSA}(\text{LN}(X_{l-1})), \quad (1)$$

$$X_l = X'_l + \text{SFFN}(\text{LN}(X'_l)), \quad (2)$$

where X'_l and X_l represent the outputs of the CSSA and the SFFN, respectively. Here, LN denotes Layer Normalization.

The algorithm's reconstruction result is obtained by the following formula: $O = F(R) + R$, where $F(\cdot)$ represents the output of the entire network, and the network is trained by minimizing the loss function:

$$\mathcal{L} = \|O - T\|_1 \quad (3)$$

where R represents the input rainy image, O represents the output clear image, and T represents the ground truth. The non-local feature extraction module mainly consists of a sparse self-attention network and a sparse feedforward neural network, which will be introduced separately in the following text.

3.2. Channel-Wise Sparse Self-Attention (CSSA)

For the sparse self-attention network, the structure is depicted in the accompanying figure. The initial features $X^1 \in \mathbb{R}^{H \times W \times C}$ obtained from the normalization layer are first expanded in dimensionality using a convolutional layer with a specified kernel size. Here, H , W , and C represent the height, width, and channel dimensions, respectively. This convolutional layer enhances the feature representation, making it more suitable for subsequent processing steps. Then, dilated convolutions with different dilation rates of 3, 2, and 1 are used to generate the Query ($Q \in \mathbb{R}^{H \times W \times C}$), Key ($K \in \mathbb{R}^{H \times W \times C}$), and Value (V) vectors, respectively [25,26]. This enables information acquisition from a broader pixel range while avoiding the gridding effect. These vectors are essential for the self-attention mechanism. The Q represents the features that need to be focused on, the K is used to measure similarities against the Q , and the V holds the information to be aggregated based on these similarities. We then compute the attention values P for all pixel pairs between the Q and K using the formula:

$$P = \frac{QK^T}{\sqrt{d}} \quad (4)$$

where $d = C/k$ is the head dimension and k is the head number. In terms of sparse self-attention, a simple but effective masking function $M(\cdot)$ is performed upon P to select the Top- k largest elements from each row of the similarity matrix. For other elements that are smaller than the threshold, we replace them with 0. This step can further filter out noisy (irrelevant) tokens and speed up the training process, which is calculated by

$$\mathcal{M}(P, k)_{i,j} = \begin{cases} P_{i,j}, & \text{if } P_{i,j} \geq \text{threshold} \\ 0, & \text{if } P_{i,j} < \text{threshold}' \end{cases} \quad (5)$$

where the threshold is a k^{th} largest value of a row. We rank the similarity scores between the query and key vectors and retain only the top k highest values, effectively filtering out irrelevant features. For the entire image, the overall complexity is reduced to $O(N \times k)$, where $k \ll N$. This computational complexity is significantly lower than the traditional dense full attention of $O(N^2)$, enabling substantial improvements in computational efficiency. The final attention matrix obtained can be represented as follows:

$$\text{Attention} = \text{softmax}(\mathcal{M}(p, k))V. \quad (6)$$

where $\text{softmax}(\cdot)$ denotes the softmax normalization function, and N represents the total number of spatial positions.

Sparse non-local self-attention is more effective than dense self-attention in single-image deraining because it selectively focuses on the most relevant features, thereby reducing the integration of irrelevant information that can blur image details.

3.3. Sparse Feedforward Network (SFFN)

In addition to its powerful self-attention mechanisms, the Transformer also incorporates a feedforward network to further enhance feature representations. To this end, we have developed a novel SFFN within the Transformer architecture tailored specifically for image deraining tasks. The structure of our SFFN is illustrated in the accompanying figure.

Traditionally, feedforward networks in Transformer rely heavily on dilated convolutions to process feature representations. However, our approach diverges by introducing dilated convolutions at varying rates (set to 2 and 3, respectively) [25,26]. This modification significantly enlarges the receptive field of the network, enabling it to capture a broader range of contextual information. The spatial sparsity introduced by dilated convolutions enriches each output position with abundant contextual information. This is especially beneficial for tasks such as image deraining, which require modeling long-range dependencies. By using dilated convolutions with different rates, our SFFN can capture both fine and coarse details in the same layer. This improves the network's ability to handle various rain patterns and densities.

Furthermore, inspired by recent advancements in the field [15], we have integrated a gating mechanism within our SFFN. This mechanism plays a critical role in dynamically balancing the flow of information between two branches of the network. By doing so, it ensures that each layer of the SFFN can concentrate on fine-grained details that supplement those identified by other layers. This complementary focus is essential for achieving high-quality image restoration, particularly in complex scenarios where intricate detail recovery is paramount. The gating mechanism dynamically adjusts the contribution of each branch according to the input features. It allows the network to adaptively prioritize different types of information as required.

Our design choices, including the variable rate dilated convolutions and the dynamic gating mechanism, collectively contribute to a more robust and efficient feedforward network. These innovations not only enhance the network's ability to capture relevant features but also reduce computational overhead by focusing on the most critical aspects of the image. This, in turn, enhances the overall performance of the Transformer in removing rain from images, providing clearer and more accurate restoration results. By enhancing both the breadth and depth of feature representation, our SFFN ensures the Transformer delivers superior deraining performance. This makes our SFFN a valuable advancement in the field of image restoration.

4. Experiments

In this section, comprehensive image deraining experiments are conducted on common benchmark datasets. We evaluate the effectiveness of the proposed method against 14 state-of-the-art image deraining baselines: DerainNet [27], SEMI [28], DIDMDN [29], UMRL [30], RESCAN [31], PReNet [32], MSPFN [33], MPRNet [34], DGUNet [35], KiT [36], Uformer-B [37], IDT [38], FADformer [39], and CSUD [40].

4.1. Datasets and Metrics

We conduct extensive experiments on the large-scale Rain13K dataset, which comprises 13,711 pairs of training images specifically curated for single-image deraining tasks. For evaluation, we utilize five widely recognized benchmark datasets: Test100 [41], Rain100H [42], Rain100L [42], Test2800 [43], and Test1200 [29]. These benchmarks cover a diverse range of rain conditions, from light drizzles to heavy downpours, ensuring a comprehensive assessment of our algorithm's effectiveness. To quantify deraining performance, we employ peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [44], two widely used metrics. PSNR measures the ratio between the maximum possible signal

power and the power of corrupting noise. It reflects the reconstruction quality in terms of pixel intensity, and higher values indicate better image quality. SSIM assesses the visual impact of differences between the restored image and the ground truth, focusing on structural information such as luminance, contrast, and texture. Both metrics are computed on the Y channel of images in the YCbCr color space. This channel is more sensitive to human visual perception and thus better reflects perceived image quality. This rigorous and objective assessment framework allows us to evaluate our proposed deraining method's performance across various challenging scenarios presented by the benchmark datasets.

4.2. Implementation Details

In our overall network architecture, the parameters {N0, N1, N2, N3, N4} are set to {2, 4, 6, 6, 8}, representing the number of Sparse Attention Transformer Blocks (SATBs) at each stage, respectively. The number of attention heads for the five Sparse Attention Transformer Blocks (SATB) at the same level is configured as {1, 1, 2, 4, 8}. The number k is set to $[1/2, 4/5]$, optimizing the balance between computational efficiency and model performance. All experiments were implemented using PyTorch 2.1.0 and trained on four NVIDIA GeForce RTX 3090 GPUs, providing a high-performance training environment. Each dataset was trained for a total of 300,000 iterations, employing the AdamW optimizer with a batch size of 16 and a patch size of 128. The initial learning rate was set to 1×10^{-4} for the first 92,000 iterations. After this stage, a cosine annealing schedule was used to adjust the learning rate progressively. This strategy improves the training process by gradually reducing the learning rate, which prevents overshooting and ensures model convergence. To further enhance the robustness of our network, we augmented the training data through horizontal and vertical flipping. This data augmentation strategy increases the diversity of the training samples, helping the model generalize better and improve its performance on unseen data. These comprehensive training settings and strategies ensure that our model is well-optimized and capable of achieving superior performance in image deraining tasks.

4.3. Comparison with State-of-the-Arts

Table 1 presents the average PSNR and SSIM metrics computed across various benchmark datasets. These metrics provide a robust quantitative assessment of the deraining performance of different models, offering insights into their efficacy in restoring image quality. PSNR measures the ratio between the maximum possible power of a signal and the power of corrupting noise, indicating how much noise has been removed and how clear the image is. SSIM assesses the similarity between two images, considering changes in structural information, luminance, and contrast. High values of PSNR and SSIM indicate better performance in terms of image restoration quality.

It is important to note that the Uformer-B, IDT, FADformer, and CSUD models were not originally trained on the Rain13K dataset [33]. To ensure a fair comparison, we retrained these models using the Rain13K dataset, aligning their training conditions with those of our proposed method. This step was essential to ensure that all models were evaluated under consistent conditions, thereby providing a more accurate comparison of their capabilities. By retraining these models on the same dataset, we eliminated potential biases and ensured that the performance differences observed were due to the models' architectures and not the training data.

As shown in Table 1, our method achieves competitive results on both PSNR and SSIM evaluation metrics. These high scores demonstrate that our model achieves excellent performance in objective image quality assessment metrics. It outperforms other methods in this field, such as the recently proposed FADformer and CSUD. Moreover, we compare our proposed sparse attention mechanism with the continuous attention method used by

IDT, and both the visual results and PSNR scores confirm the effectiveness of our approach. To further validate the efficiency of the proposed method, we select several competitive approaches from Table 1. These approaches include DGUNet, KiT, Uformer-B, FADformer, and CSUD, and we perform a comprehensive complexity analysis and comparison. As shown in Table 2, our model achieves the best balance between performance and computational cost. Specifically, our method requires only 11.6M parameters and 31.20G FLOPs, which are the lowest among all competing approaches. This also indirectly explains why some metrics of DGUNet and KiT are higher than ours—the marginal improvements are essentially traded for complex model architectures. Our model consumes less inference time and memory usage, making it more suitable for deployment on resource-constrained devices. These results demonstrate that our sparse attention mechanism not only maintains competitive image restoration quality but also significantly improves computational efficiency. It provides a more practical solution for real-world image deraining applications.

Table 1. Comparison of quantitative results on five benchmark datasets. Bold indicates the best results.

Datasets	Test100		Rain100H		Rain100L		Test2800		Test1200		Average	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DerainNet	22.77	0.810	14.92	0.592	27.03	0.884	24.31	0.861	23.38	0.835	22.48	0.796
SEMI	22.35	0.788	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822	22.88	0.744
DIDMDN	22.56	0.818	17.35	0.524	25.23	0.741	28.13	0.867	29.95	0.901	24.64	0.770
UMRL	24.41	0.829	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.910	28.02	0.880
RESCAN	25.00	0.835	26.36	0.786	29.80	0.881	31.29	0.904	30.51	0.882	28.59	0.858
PReNet	24.81	0.851	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911	29.43	0.897
MSPFN	27.50	0.876	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	30.75	0.903
MPRNet	30.27	0.897	30.41	0.890	36.40	0.965	33.64	0.938	32.91	0.916	32.73	0.921
DGUNet	30.32	0.899	30.66	0.891	37.42	0.969	33.68	0.938	33.23	0.920	33.06	0.923
KiT	30.26	0.904	30.47	0.897	36.65	0.969	33.85	0.941	32.81	0.918	32.81	0.926
Uformer-B	29.90	0.906	30.31	0.900	36.86	0.972	33.53	0.939	29.45	0.903	32.01	0.924
IDT	29.69	0.905	29.95	0.898	37.01	0.971	33.38	0.937	31.38	0.908	32.28	0.924
FADformer	30.05	0.899	30.33	0.901	37.11	0.960	33.62	0.934	31.56	0.907	32.53	0.920
CSUD	31.11	0.907	24.42	0.808	33.28	0.954	33.83	0.940	31.28	0.901	30.78	0.902
Ours	31.53	0.916	30.14	0.885	37.52	0.970	33.94	0.941	32.13	0.916	33.05	0.926

Table 2. Performance comparison of different models.

Models	DGUNet	KiT	Uformer-B	FAD-Former	CSUD	Ours
Params (M)	17.33	20.6	50.88	16.95	12.38	11.6
FLOPs (G)	199.87	43.08	89.46	48.51	150.03	31.20
Inference Time (s)	0.181	0.132	0.169	0.146	0.083	0.019
Memory Usage (GB)	0.556	0.597	0.667	0.528	0.383	0.355

Furthermore, we conduct visual comparisons to intuitively demonstrate the deraining effects of different models. Figures 2 and 3 present the visualization results on the Rain100H and Test100 datasets, respectively. These figures clearly demonstrate the effectiveness of our model in removing rain streaks and preserving fine image details. Images processed by our model exhibit noticeably fewer rain streaks and retain more intricate details, such as textures and edges, which are crucial for high-quality image restoration. Additionally, to further demonstrate the generalization capability of our model, we compare it with competing methods on real-world rainy images. As shown in Figure 4, our model exhibits certain advantages in both overall image clarity and detailed texture preservation when examining zoomed-in regions.

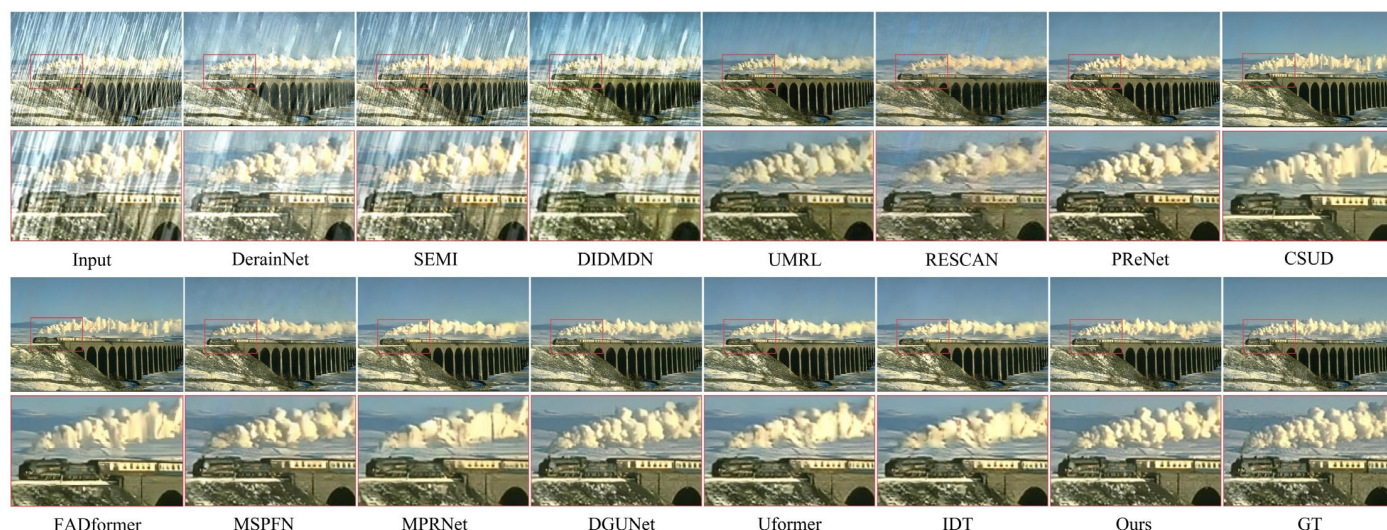


Figure 2. Visual quality comparison of deraining images obtained by different methods on the Rain100H benchmark dataset. The red box indicates the zoomed-in effect.



Figure 3. Visual quality comparison of deraining images obtained by different methods on the Test100 benchmark dataset. The red box indicates the zoomed-in effect.



Figure 4. Visual quality comparison of deraining images obtained by different methods on real-world rainy images. The red box indicates the zoomed-in effect.

The images processed by our model exhibit significantly better rain removal and detail preservation. This superior performance can be attributed to the innovative design and the inherent advantages of our developed framework. Our framework effectively integrates sparse self-attention and a sparse feedforward network, which are crucial components for achieving high performance. Sparse self-attention allows the model to focus on the most relevant features by filtering out noise and irrelevant information, enhancing the clarity and accuracy of the restored images. Meanwhile, the sparse feedforward network ensures that detailed spatial and channel-wise information is preserved and utilized effectively during the deraining process. These components enable more accurate and efficient feature aggregation, ultimately leading to clearer and more visually appealing restoration outcomes. This demonstrates the practical effectiveness of our approach in enhancing image quality and providing more realistic and high-fidelity results. The combination of these innovative techniques validates our model's robustness and effectiveness, setting a new standard in the field of image deraining. Our approach improves not only quantitative metrics but also the qualitative performance of images. It significantly enhances visual appeal and detail retention, making it a comprehensive solution for image deraining tasks.

4.4. Ablation Studies

We conducted an in-depth study to analyze the impact of the main components of our proposed method on the final deraining performance. Specifically, we considered the following variants of our method: (a) Baseline model (without Top-k and dilated convolutions); (b) model with dilated convolutions only; (c) Model with Top-k sparse attention only; (d) our proposed integrated model (with both Top-k sparse attention and dilated convolutions). These variations allow us to isolate and evaluate the contribution of each component to the overall performance of our model. Table 3 presents the quantitative evaluation results of these different configurations on the Rain100H benchmark dataset.

Table 3. Ablation study analysis on the Rain100H benchmark dataset. Bold indicates the best results.

Models	(a)	(b)	(c)	(d)
baseline	✓	✓	✓	✓
Top-k			✓	✓
Dailated Conv		✓		✓
PSNR	28.57	29.79	29.87	30.14
SSIM	0.859	0.867	0.870	0.885

Our analysis indicates that model (d), which includes Top-k sparse attention and dilated convolutions, consistently outperforms other configurations. This superior performance underscores the effectiveness of each individual component in enhancing the deraining capability of our network. The inclusion of the Top-k sparse attention mechanism significantly improves the model's ability to filter out irrelevant features, leading to clearer and more accurate image restoration. By focusing on the most relevant similarities, the Top-k sparse attention mechanism enhances the network's efficiency and effectiveness. Overall, the results demonstrate that each of these components plays a crucial role in boosting the performance of our deraining model. The combination of these innovations brings significant improvements to both PSNR and SSIM metrics. This validates the robustness and effectiveness of our proposed approach. This comprehensive evaluation highlights the importance of each component and confirms that our method's design choices significantly contribute to its superior deraining performance.

Effect of the number of k. We observe that the optimal selection of k plays a crucial role in controlling the sparsity rate within the model. When k is manually fixed to a specific

value, such as $1/2$, we found that the model's performance is highly sensitive to this choice. To circumvent the need for an exhaustive search, we introduce a controllable interval for k that allows the model to dynamically learn the most contributive score. Our approach achieves the best performance, reaching 31.53 dB, when the interval $[\Delta 1, \Delta 2]$ is assigned to $[1/2, 4/5]$.

5. Limitations

Despite the promising results of our proposed image deraining method, several limitations remain. First, the computational demands of our Transformer-based approach, although reduced through sparse attention mechanisms, still pose challenges for deployment on devices with limited processing power. Second, while our method is specifically designed for rain removal, it may not effectively address other weather-related artifacts such as fog or snow. This indicates a need for further generalization and adaptation of our approach to handle a broader range of adverse weather conditions, enhancing its overall robustness and applicability in diverse scenarios.

6. Conclusions

In this work, we have proposed an effective image deraining Transformer leveraging sparse non-local self-attention mechanisms. Our approach replaces the conventional dense self-attention with a novel sparse self-attention strategy, enhancing the model's ability to focus on the most relevant features and improving the rain removal process. Additionally, we integrated a sparse feed-forward network to refine feature representations, ensuring more accurate and efficient image restoration. Extensive quantitative and qualitative experiments on various rain streak benchmarks show that our model significantly outperforms state-of-the-art methods in both PSNR and SSIM. This demonstrates its superior ability to restore high-quality, rain-free images. Real-world scenarios often involve multiple co-existing degradations, presenting greater challenges. In the future, we will investigate whether different image degradations (rain, blur, noise, etc.) can be modeled within a unified image restoration framework, enabling a single model to handle various mixed degradation problems.

Author Contributions: Conceptualization, methodology, framework, implementation, validation, and writing—original paper draft, X.Z.; supervision, direction and planning, review, funding acquisition, resources, writing—review and editing, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Shenyang Science and Technology Project under Grant 23-503-6-18, the Fundamental Research Funds for the Universities of Liaoning Province under Grant LJ232410143060, the Scientific Research Platform Construction Project of the Education Department of Liaoning Province under Grant LJ232510143007, and the Liaoning Provincial Applied Basic Research Project under Grant 2025JH2/101330120.

Data Availability Statement: The datasets used in this study are publicly accessible.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kang, L.W.; Lin, C.W.; Fu, Y.H. Automatic single-image-based rain streaks removal via image decomposition. *IEEE Trans. Image Process.* **2011**, *21*, 1742–1755. [[CrossRef](#)] [[PubMed](#)]
2. Li, Y.; Tan, R.T.; Guo, X.; Lu, J.; Brown, M.S. Rain streak removal using layer priors. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 2736–2744.

3. Yu, Y.; Yang, W.; Tan, Y.P.; Kot, A.C. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 6013–6022.
4. Fan, Z.; Wu, X.; Chen, X.; Li, Y. Learning to see in nighttime driving scenes with interfrequency priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2023; pp. 4217–4224.
5. Chen, X.; Pan, J.; Dong, J. Bidirectional multi-scale implicit neural representations for image deraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 25627–25636.
6. Chai, G.; Yang, R.; Ge, J.; Chen, Y. A Hybrid Content-Aware Network for Single Image Deraining. *Computers* **2025**, *14*, 262. [\[CrossRef\]](#)
7. Chen, H.; Chen, X.; Lu, J.; Li, Y. Rethinking multi-scale representations in deep deraining transformer. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2024; Volume 38, pp. 1046–1053.
8. Wang, L.; Gao, G. DECTNet: A detail enhanced CNN-Transformer network for single-image deraining. *Cogn. Robot.* **2025**, *5*, 48–60. [\[CrossRef\]](#)
9. Chen, X.; Pan, J.; Jiang, K.; Li, Y.; Huang, Y.; Kong, C.; Dai, L.; Fan, Z. Unpaired deep image deraining using dual contrastive learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 2017–2026.
10. Jiang, K.; Jiang, J.; Liu, X.; Xu, X.; Ma, X. FMRNet: Image deraining via frequency mutual revision. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2024; Volume 38, pp. 12892–12900.
11. Li, Y.; Fan, Z.; Lu, J.; Chen, X. Tao-net: Task-adaptive operation network for image restoration and enhancement. *IEEE Signal Process. Lett.* **2022**, *29*, 2198–2202. [\[CrossRef\]](#)
12. Xu, N.; Yang, J.J. Leveraging Scene Geometry and Depth Information for Robust Image Deraining. *Computers* **2025**, *14*, 11. [\[CrossRef\]](#)
13. Zou, W.; Wang, Y.; Fu, X.; Cao, Y. Dreaming to prune image deraining networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 6023–6032.
14. Yao, Y.; Shi, Z.; Hu, H.; Li, J.; Wang, G.; Liu, L. GSDerainNet: A deep network architecture based on a Gaussian shannon filter for single image deraining. *Remote Sens.* **2023**, *15*, 4825. [\[CrossRef\]](#)
15. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 5728–5739.
16. Zhou, H.; Chen, H.; Wu, X.; Li, Y. Dual-path multi-scale transformer for high-quality image deraining. In *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
17. Wu, X.; Chen, H.; Chen, X.; Xu, G. Multi-scale transformer with conditioned prompt for image deraining. *Digit. Signal Process.* **2025**, *156*, 104847. [\[CrossRef\]](#)
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
19. Zhao, H.; Jia, J.; Koltun, V. Exploring self-attention for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 10076–10085.
20. Gautam, A.; Pawar, A.; Joshi, A.; Tazi, S.N.; Chaudhary, S.; Hambarde, P.; Dudhane, A.; Vipparthi, S.; Murala, S. Pureformer: Transformer-Based Image Denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 1441–1449.
21. Cao, X.; Wang, X.; Dun, X.; Lian, Y.; Cheng, X.; Hao, X. Cross-domain-aware deep unfolding transformer for hyperspectral image super-resolution. *Pattern Recognit.* **2025**, *172*, 112374. [\[CrossRef\]](#)
22. Xu, S.; Sun, Z.; Zhong, M.; Cao, C.; Liu, Y.; Fu, X.; Chen, Y. Motion-adaptive Transformer for Event-based Image Deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2025; Volume 39, pp. 8942–8950.
23. Hu, T.; Chen, Z.; Wang, M.; Hou, X.; Lu, X.; Pan, Y.; Li, J. Global sparse attention network for remote sensing image super-resolution. *Knowl.-Based Syst.* **2024**, *304*, 112448. [\[CrossRef\]](#)
24. Gendy, G.; Sabor, N. Blind image super-resolution using swin transformer with unsupervised degradation and sparse attention. *Neural Comput. Appl.* **2025**, *37*, 21493–21517. [\[CrossRef\]](#)
25. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
26. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2018; pp. 1451–1460.
27. Fu, X.; Huang, J.; Ding, X.; Liao, Y.; Paisley, J. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Trans. Image Process.* **2017**, *26*, 2944–2956. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Wei, W.; Meng, D.; Zhao, Q.; Xu, Z.; Wu, Y. Semi-supervised transfer learning for image rain removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 3877–3886.

29. Zhang, H.; Patel, V.M. Density-aware single image de-raining using a multi-stream dense network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 695–704.
30. Yasarla, R.; Patel, V.M. Uncertainty guided multi-scale residual learning-using a cycle spinning CNN for single image de-raining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 8405–8414.
31. Li, X.; Wu, J.; Lin, Z.; Liu, H.; Zha, H. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 254–269.
32. Ren, D.; Zuo, W.; Hu, Q.; Zhu, P.; Meng, D. Progressive image deraining networks: A better and simpler baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 3937–3946.
33. Kui 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 8346–8355.
34. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Multi-stage progressive image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 14821–14831.
35. Mou, C.; Wang, Q.; Zhang, J. Deep generalized unfolding networks for image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 17399–17410.
36. Lee, H.; Choi, H.; Sohn, K.; Min, D. KNN local attention for image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 2139–2149.
37. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 17683–17693.
38. Xiao, J.; Fu, X.; Liu, A.; Wu, F.; Zha, Z.J. Image de-raining transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 12978–12995. [[CrossRef](#)] [[PubMed](#)]
39. Gao, N.; Jiang, X.; Zhang, X.; Deng, Y. Efficient frequency-domain image deraining with contrastive regularization. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 240–257.
40. Dong, G.; Zheng, T.; Cao, Y.; Qing, L.; Ren, C. Channel Consistency Prior and Self-Reconstruction Strategy Based Unsupervised Image Deraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 7469–7479.
41. Zhang, H.; Sindagi, V.; Patel, V.M. Image de-raining using a conditional generative adversarial network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3943–3956. [[CrossRef](#)]
42. Yang, W.; Tan, R.T.; Feng, J.; Liu, J.; Guo, Z.; Yan, S. Deep joint rain detection and removal from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 1357–1366.
43. Fu, X.; Huang, J.; Zeng, D.; Huang, Y.; Ding, X.; Paisley, J. Removing rain from single images via a deep detail network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 3855–3863.
44. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.