Electronic
Research Archive

https://www.aimspress.com/journal/era

ERA, 33(8): 4740–4762.
DOI: 10.3934/era.2025213
Received: 22 June 2025
Revised: 18 July 2025
Accepted: 04 August 2025
Published: 15 August 2025

Research article

E^2-ViTUNet: Enhanced dual-encoder vision transformer UNet for single-image rain removal

Huirong Fang[1,†], Zhixiang Chen[2,†,*], Ziyang Zheng[2] and Hui Wang[2]

[1] School of Electronic Information, Zhangzhou Institute of Technology, Zhangzhou 363000, China

[2] School of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China

† The authors contributed equally to this work.

* **Correspondence:** Email: zxchenphd@163.com.

Abstract: The rapid development of image deraining technology has made it possible to effectively restore images captured even in severe rainy weather, thereby alleviating the decline in image quality. However, existing image deraining solutions typically use overly deep neural networks, which are prone to gradient vanishing and explosion. This often results in the restored image losing essential background details. To address this important issue, this paper proposed an enhanced dual-encoder vision transformer UNet for single-image rain removal. First, a dual-branch heterogeneous encoding structure was formed by integrating two encoders with different architectures. These encoders extract rain streaks by leveraging their distinct characteristics. Specifically, the first encoder was based on a convolutional neural network, termed the residual encoder network, which utilizes the local receptive field capabilities of convolutional kernels to capture both rain streaks and corresponding background details. Concurrently, the second encoder, the parallel vision transformer encoder network, employs self-attention mechanisms to model long-range dependencies and establish global contextual relationships between rain streaks and background details. Subsequently, the rain streak information extracted by the two encoders was fused. The decoder then performs the rain removal task. To preserve spatial consistency and overall image integrity, high-frequency detail refinement was facilitated through residual connections between the encoders and the decoder. Evaluations on synthetic and real-world datasets confirm the algorithm's robustness in processing rain streaks of varied density. The restored images exhibit enhanced visual clarity, improving visibility for computer vision applications.

Keywords: rain removal; vision transformer; neural network; encoder; attention mechanism

1. Introduction

Image deraining technology aims to remove raindrops, rain streaks, and rain-induced degradations from rain-contaminated visual data to reconstruct perceptually clear content. This technology is particularly vital in critical application domains including autonomous driving, video surveillance, and computational photography. As shown in Figure 1, rain perturbations severely compromise visual clarity and structural fidelity, consequently degrading the robustness of downstream computer vision algorithms. However, existing deraining techniques predominantly target narrow quantitative metrics, relying on over-parameterized network architectures while overlooking cross-task synergy with higher-level vision applications.

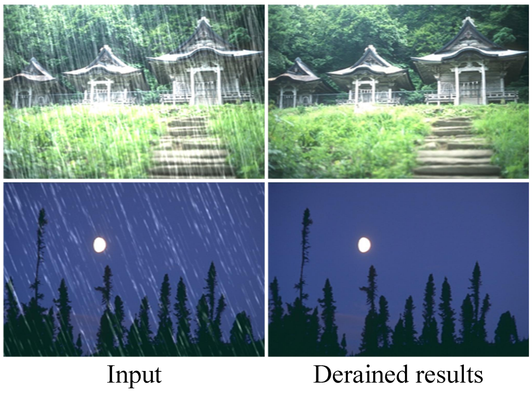


Figure 1. Sample results of the proposed E^2-ViTUNet method for single-image deraining.

Image degradation in rain and fog scenes primarily manifests as raindrops of varying sizes forming rain streaks distributed across the entire image plane. These rain streaks induce localized dense degradation on imaging. From a mathematical perspective, a rainy image can be decomposed into two underlying constituent layers: The weather-induced interference layer representing rain streaks, and the latent scene content layer retaining intrinsic scene structures. Therefore, the decomposition problem is expressed as:

$$O = R + B \tag{1.1}$$

where O represents the rainy day image, R represents the rain streaks, and B represents the clean background image (Figure 2). Therefore, the core objective of single-image deraining is to accurately segregate the two constituent layers via optimized algorithmic frameworks [1,2].

However, it remains challenging to preserve high-frequency background details while optimizing network architecture and reducing model depth. Current deraining solutions are broadly categorized into two paradigms: prior-based methods (traditional optimization-based approaches) and data-driven methods (deep neural network-based approaches). Prior-based approaches employ mathematical formulations to disentangle clean backgrounds from rainy images by exploiting intrinsic properties of natural scenes and rain streaks (e.g., background smoothness and streak sparsity). Although existing frameworks incorporate prior analysis of natural images [3,4] with techniques including Gaussian mixture modeling (GMM) [5], sparse convolutional coding [3], and multimodal low-rank analysis [6, 7], attempting to establish a layer decomposition model through regularization-driven priors. Real-world rainy scenes exhibit intricate complexities. These approaches often induce artificial truncation of high-frequency components, critically compromising artifact suppression capabilities.

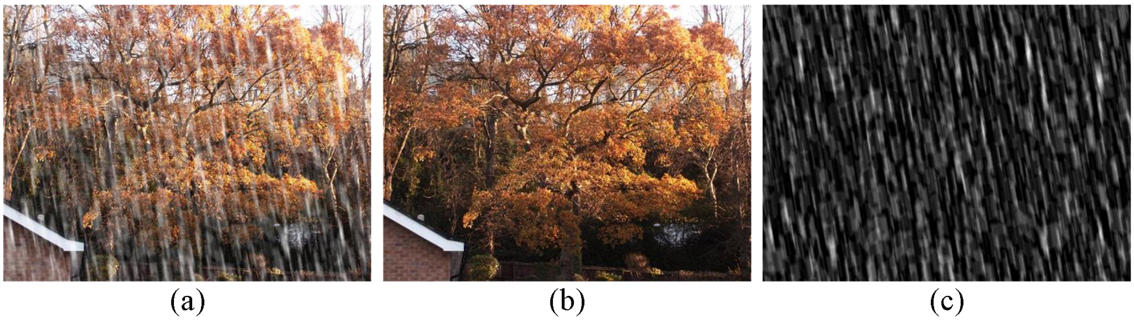


Figure 2. Single-image rain streak removal. Conceptually, a rain-affected input image (a) comprises the additive superposition of the latent scene structure (b) and the sparsely distributed rain layer (c).

In recent years, the computer vision field has witnessed breakthrough advances in image deraining technology [8, 9], where deep learning frameworks [10, 11] have pioneered new avenues for image restoration under complex conditions. Within this paradigm, data-driven methods [5, 12] demonstrate superior efficacy in addressing this notoriously ill-posed problem. End-to-end convolutional neural network (CNN) architectures [11, 13] have overcome fundamental bottlenecks in feature representation capacity and model generalization inherent in traditional prior-based models [4, 5]. Furthermore, researchers have designed specialized architectural components like visual transformers [14], encoder-decoder structures [15], residual connections [16], and skip connections [17] to optimize performance. Nevertheless, these deraining architectures remain susceptible to unstable gradient propagation.

Unstable gradient propagation occurs when derivatives cause gradients to decay/grow exponentially fast during backpropagation through the chain rule. This phenomenon impedes the effective updating of early-layer parameters (vanishing) or drives them to overshoot catastrophically (explosion). Ignoring this issue induces training failure in deep networks, causing model divergence/oscillation and parameter invalidation, resulting in computational resource wastage and degraded restoration fidelity. Conversely, if this pathology is successfully mitigated, it enables the construction of architecturally deeper networks, improves model convergence efficiency and training stability, and unlocks the full modeling capacity.

To overcome these limitations, we introduce an enhanced dual-encoder vision transformer UNet (E^2-ViTUNet) for single-image deraining. The key technical contributions are summarized as follows.

1) We design a dual-stream feature extraction architecture combining a convolutional encoder (capturing local rain streak details) with an innovative parallel self-attention encoder (modeling global context). This enables complementary local-global representations, enhancing discrimination capacity for rain-affected regions while preserving the structural integrity of background details.

2) The restructured vision transformer block employs a parallel architecture that expands feature width horizontally while reducing layer stacking depth. This configuration simultaneously lowers computational complexity and enhances hierarchical global modeling capabilities for multi-scale rain streaks, substantially improving processing throughput for high-resolution images.

3) Dual-encoded feature aggregation and cross-level residual connections achieve multi-level feature fusion and stabilize decoder gradients. Mitigating gradient instability and enhancing shallow feature propagation optimizes degradation removal while preserving original scene textures.

4) Extensive experiments confirm the efficacy of E^2-ViTUNet on image deraining. The results demonstrate its superiority over advanced methods, attaining near state-of-the-art performance. Additionally, E^2-ViTUNet shows excellent generalization in downstream vision tasks.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 states the problem. Section 4 introduces the proposed framework. Section 5 presents the evaluation results. Section 6 discusses applicability issues. Section 7 demonstrates industrial applications. Finally, Section 8 concludes the paper.

2. Related work

Research progress in single-image rain removal algorithms has mainly focused on two core technical approaches: prior-based methods and deep-learning-based methods. These two rain removal paradigms exhibit significant differences in their theoretical frameworks and implementation paths. In this section, we will provide a systematic overview and analysis of their core principles and typical methods.

2.1. Prior-based methods

single-image rain removal is a typical underdetermined inverse problem, and its core challenge lies in how to stably reconstruct the underlying clean scene from degraded observations. Early research mainly used a variational framework, which designed heuristic sparse constraints or texture covariance priors to shrink the non-unique solution space of the inverse problem to a reasonable signal manifold, thus ensuring the numerical stability of iterative optimization. Luo et al. [4] developed a DSC-based method for layer-decoupled rain decomposition, effectively preserving image semantics during precipitation suppression. Li et al. [5] proposed an image decomposition method (GMM) based on hierarchical priors (such as brightness, directionality, etc.). By modeling the physical differences between the rain streak layer and the background layer, rain streaks are separated and removed from a single image. Ran et al. [12] introduced directional gradient features to enhance rain streak detection capabilities, combined multi-scale fusion and residual learning mechanisms, and effectively utilized the directional characteristics of rain streaks to remove the streaks from single images (DiG-CoM). The above-mentioned a priori rain removal methods can remove some rain streaks, but the modeling process is complex and it is difficult to remove heavy rain streaks.

2.2. Deep-learning-based methods

Currently, image deraining algorithms based on deep neural networks have achieved significant results in the field of computer vision. In particular, deep learning methods represented by convolutional neural networks have demonstrated breakthrough performance in restoring visual details blurred by rain. By constructing end-to-end feature learning frameworks, which effectively achieve accurate modeling and removal of rain streaks, Fu et al. [18] effectively solved the problem of rain streak degradation and background detail confusion (DDN) by decomposing the joint learning framework of the base layer and detail layer of the image. Li et al. [19] proposed a recurrent squeeze-and-aggregate context aggregation network (RESCAN), which refines rain streak features step by step through a recurrent structure and combines a context aggregation module (integrating global and local information) with a channel attention mechanism to adaptively enhance the representation ability of background details in complex rain streak scenes. Ren et al. [20] used a multi-stage progressive optimization strategy and a simplified network architecture (PReNet) to gradually separate rain streaks from background textures during iterative rain removal, significantly improving rain removal performance and model robustness. Wang et al. [21] achieved more accurate rain streak separation and background restoration (RCDNet) by embedding physical priors (such as sparsity and directionality constraints of rain streaks) into the network architecture and jointly learning the image decomposition and reconstruction processes. Zamir et al. [13] proposed a multi-stage progressive image restoration framework (MPRNet) that gradually reconstructs high-quality images from low-quality inputs at multiple degradation levels, combining global structure restoration with local detail enhancement to significantly improve restoration performance in complex degradation scenarios. Zhang et al. [11] proposed the enhanced spatio-temporal interaction network (ESTINet), which achieves efficient spatio-temporal feature coupling and redundant computation reduction in video deraining tasks through dynamic sparse convolution strategies and cross-frame feature adaptive aggregation mechanisms, significantly improving rain streak removal accuracy while ensuring real-time performance. While existing algorithm frameworks have made breakthrough progress, they still face key technical bottlenecks. First, there is still no effective mechanism to suppress the phenomenon of gradient disappearance and explosion caused by increased network depth. Second, when facing downstream tasks, there is no coordinated optimization mechanism with high-level semantic understanding (such as semantic segmentation or object detection). This directly limits the application potential of this technical approach in complex scenarios.

3. Proposed formulation

In this section, we address the issue of gradient vanishing/explosion in network training within deraining scenarios and its impact on suboptimal detail restoration fidelity. In such contexts, traditional single-path encoder-decoder networks face a dilemma: If deep architectures are employed to capture global features, gradient attenuation during backpropagation intensifies exponentially with network depth. If shallow architectures are adopted to suppress gradient attenuation, the limited receptive field results in insufficient modeling capacity for high-frequency details. Critically, the spectral overlap between rain streaks and background textures induces gradient competition. When rain streaks dominate the propagation process, it may cause gradient explosion; conversely, it may induce gradient vanishing. Therefore, a dual-path heterogeneous encoding structure is required to circumvent gradient vanishing/explosion while enhancing rainy image restoration performance.

Given the analogous causes of gradient vanishing/explosion in CNNs [22] and recurrent neural networks (RNNs) [23], this paper employs RNNs as an analytical exemplar [24]. In recurrent neural network training, this phenomenon is attributed to the multiplicative effect of weight matrices over extended sequences. Specifically, we consider the hidden state update in a basic RNN:

$$h_t = \sigma\left(W \cdot h_{t-1} + U \cdot x_t + b\right) \tag{3.1}$$

where h_t represents the hidden layer output vector, σ represents the non-linear activation function, W represents the recurrent weight matrix, U represents the input weight matrix, x_t represents the input vector, and b represents the bias vector.

During backpropagation, the gradient of the loss function with respect to the initial hidden state h_0 is calculated as:

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_T} \cdot \prod_{k=1}^{T} \frac{\partial h_k}{\partial h_{k-1}} = \frac{\partial L}{\partial h_T} \cdot \prod_{k=1}^{T} \operatorname{diag}\left(\sigma'\left(W h_{k-1} + U x_k\right)\right) \cdot W \tag{3.2}$$

where $\frac{\partial L}{\partial h_0}$ represents the gradient of the loss function L with respect to the initial h_0, T represents the total length of the time series, $\frac{\partial h_k}{\partial h_{k-1}}$ represents the Jacobian matrix of h_k with respect to h_{k-1}, σ' represents the derivative of the activation function σ, and diag$(\cdot)$ represents the conversion of a vector to a diagonal matrix.

Specifically, gradient vanishing causes the network to lose its ability to evolve features across layers. Gradient propagation in deep encoding-decoding structures also satisfies the matrix chain rule. Let the network have L layers, with each layer's transformation matrix being J_l (corresponding to the product of the derivative of the activation function and the weight matrix). Then, gradient propagation can be modelled as:

$$\frac{\partial L}{\partial \theta_l} = \frac{\partial L}{\partial y_L} \cdot \prod_{k=L}^{l+1} J_k \tag{3.3}$$

where θ_l represents the learnable parameters of layer l, y_L represents the output of layer L (the last layer) of the network, and J_k represents the Jacobian matrix of layer k.

Taking the traditional single-path encoder-decoder architecture as an example, the feature map contraction caused by continuous downsampling significantly diminishes the response intensity of shallow convolutional kernels, causing the extracted rain streaks to degenerate into noise signals during propagation to the decoder. This ultimately prevents the reconstruction module from accurately localizing rain streak boundaries, resulting in blurred rain streak artifacts in the deraining image. Conversely, gradient explosion can trigger feature space collapse. When the singular value distribution of a weight matrix at a certain layer substantially deviates from equilibrium, the backpropagation error undergoes cumulative amplification along specific channels, causing the parameter update process to diverge from optimal trajectories. Specifically, this manifests as grid-shaped artifact patterns or localized color distortion in deraining results.

Different network architectures exhibit varying susceptibility to gradient vanishing/explosion. To achieve superior image restoration performance in single-image deraining, our optimization objective is to minimize the pathological manifestations of gradient vanishing/explosion, or mitigate the pathological effects while preserving structural integrity.

4. Single-image rain removal with E^2-ViTUNet

This section proposes a dual-path heterogeneous encoding framework for image deraining tasks, establishing a complementary learning mechanism for local representations and global semantics through a convolutional encoder coupled with an improved parallel self-attention encoder. A dual-coding feature stacking strategy combined with cross-level residual skip connections achieves balanced optimization between rain streak suppression and texture detail restoration in real-world scenes.

Existing image deraining methods typically suffer from overly deep architectures. Excessively deep convolutional stacks are liable to induce training instability as network depth increases; this often renders hidden-layer parameters uncontrollable, inducing gradient vanishing/explosion degradation that ultimately causes computational redundancy and limited generalization capability. To overcome these limitations, we propose a new dual-path heterogeneous encoding framework for image deraining tasks. It establishes a collaborative modeling mechanism between local detail response and global context awareness, enabling precise identification of rain streak contamination areas while maintaining structural coherence of background content and preserving textural details. In Algorithm 1, the framework implements a local-global feature complementary scheme.

Algorithm 1 Dual-Encoder Frame Algorithm

```
Require: Rain image as $Input\_Img$
Ensure: Fused feature map as $F_0$
 1: $n = 0, m = 0, F_0 = Input\_Img$
 2: while $n \leq 4$ do
 3:     $F_1 = Convolution\ Layer\ \&\ Downsampling(F_0)$
 4:     while $m \leq L$ do
 5:         $F_2 = F_1$
 6:         $F_2 = Transformer\ Block(F_2)$
 7:         $m = m + 1$
 8:     end while
 9:     $F_2 = Layer\ Norm\ \&\ MLP\ Head\ \&\ CBAM(F_2)$
10:     $F_0 = F_1 + F_2$
11:     $n = n + 1$
12: end while
```

Step 1: The rainy image (denoted by $Input\_Img$) serves as algorithm input (Line 1), processed by a five-stage convolutional encoder with 3 × 3 kernels (channel dimensions: 64 → 128 → 256 → 512 → 1024) and downsampling. This yields a feature map (denoted by F_1) at resolutions from 256 × 256 to 16 × 16 (Line 3).

Step 2: RENet's output serves as PVTENet input (denoted by F_2, Line 5). After progressing through L Transformer Blocks, LayerNorm, MLP Head, and CBAM, F_2 is refined (Lines 4–9).

Step 3: Element-wise summation of F_1 and F_2 produces fused feature map F_0, completing the first iteration. F_0 subsequently feeds into RENet for the next iteration (Line 3).

Building upon this framework, we propose E^2-ViTUNet (Figure 3). The scheme mitigates gradient vanishing/explosion while enhancing deraining performance, operating in three phases:

Step 1: Algorithm 1 processes rainy images via RENet and PVTENet modules, generating rain streak feature map F_0.

Step 2: The decoder mirrors RENet's structure, replacing $2 \times$ *downsampling* with $2 \times$ *upsampling* between convolutional layers. Residual connections propagate multi-scale information to strengthen gradient flow.

Step 3: Skip connections concatenate RENet's feature maps with corresponding decoder layers, blending low-level rain streaks with high-level semantics for feature complementarity. Finally, the decoder ultimately outputs high-resolution deraining images.

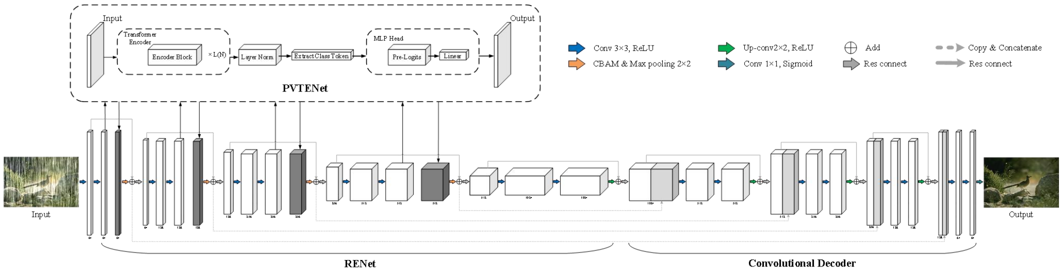


Figure 3. The structure diagram of E^2-ViTUNet.

4.1. Design of the RENet encoder

To capture local rain streak details, we construct the residual encoder network (RENet) using convolutional residual blocks (Figure 4). This architecture outperforms traditional encoders with fewer parameters, effectively mitigating representation degradation caused by deep networks.

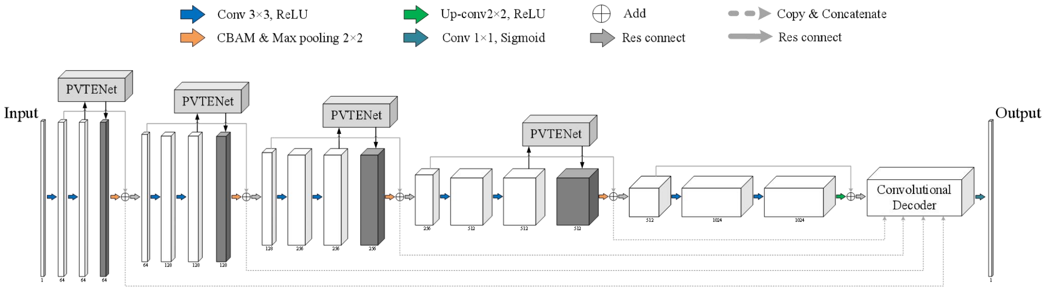


Figure 4. The structure diagram of RENet.

Accepting 256×256 RGB inputs, RENet employs a five-stage pipeline. Convolutional filters per stage progressively double ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$). Each stage comprises two 3×3 convolutional layers, a 1×1 shortcut layer, and 2×2 max pooling for downsampling, achieving feature sizes of $256 \times 256 \rightarrow 128 \times 128 \rightarrow 64 \times 64 \rightarrow 32 \times 32 \rightarrow 16 \times 16$. Layer normalization and activation functions process outputs after each block.

For improved rain streak localization, we insert the convolutional block attention module (CBAM) [25] before each downsampling, calibrating inter-channel dependencies in RGB space. As Figure 5 shows, CBAM combines two principle-attention mechanisms:

1) Channel attention module (CAM) dynamically adjusts feature channel allocation to enhance position-sensitive features related to rain streaks. Operation sequence (Figure 6): Input feature F undergoes simultaneous global max pooling and average pooling → parameter-shared MLP compression → element-wise summation → sigmoid activation → weight map M_c → channel-refined feature F'.

2) Spatial attention module (SAM) focuses on salient regions (e.g., rain streak locations). Process flow (Figure 6): F' undergoes channel-wise max/avg pooling → 7×7 convolution fusion → sigmoid activation → spatial weight M_s → spatial-refined feature F''.

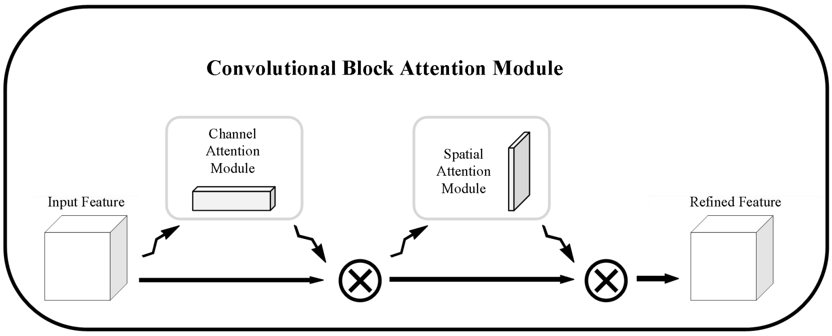


Figure 5. Overview of the convolutional block attention module (CBAM). CBAM incorporates two cascaded sub-components for channel and spatial feature refinement. At each convolutional stage within the deep network architecture, intermediate feature activations undergo attention-based recalibration via the CBAM mechanism prior to propagation.

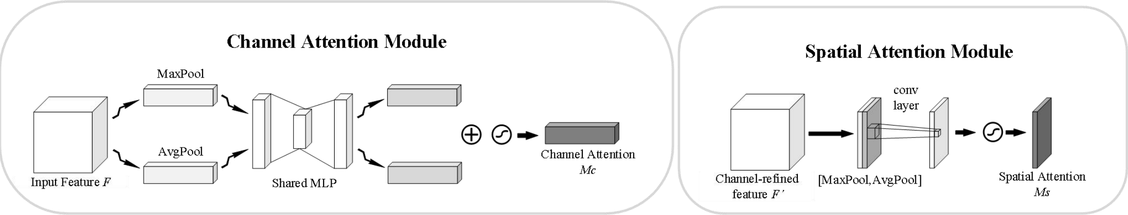


Figure 6. Detailed architecture of CBAM's attention components. The channel attention mechanism (left) incorporates a shared multilayer perceptron, processing features derived from both max-pooling and average-pooling operations across spatial dimensions. Conversely, the spatial attention module (right) employs channel-wise pooled feature statistics, subsequently encoded by a convolutional layer prior to attention map generation.

4.2. Design of the PVTENet encoder

Extending the vision transformer, we propose a parallel vision transformer encoder network (PVTENet) (Figure 7) for global rain streak feature extraction, markedly improving the distinction of rain-covered regions while preserving background detail integrity.

Originally developed for NLP tasks (e.g., machine translation) [26], the vision transformer (ViT) delivers versatility through patch-based representations. As shown in Figure 7, an input image is processed through L identical transformer encoder blocks, enabling mutual attention among image patches. After layer normalization, the class token is extracted as the global feature representation, progressed through Pre-Logits and Linear layers within the MLP Head for feature refinement.

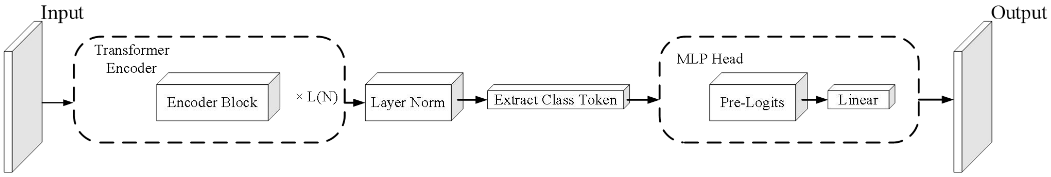
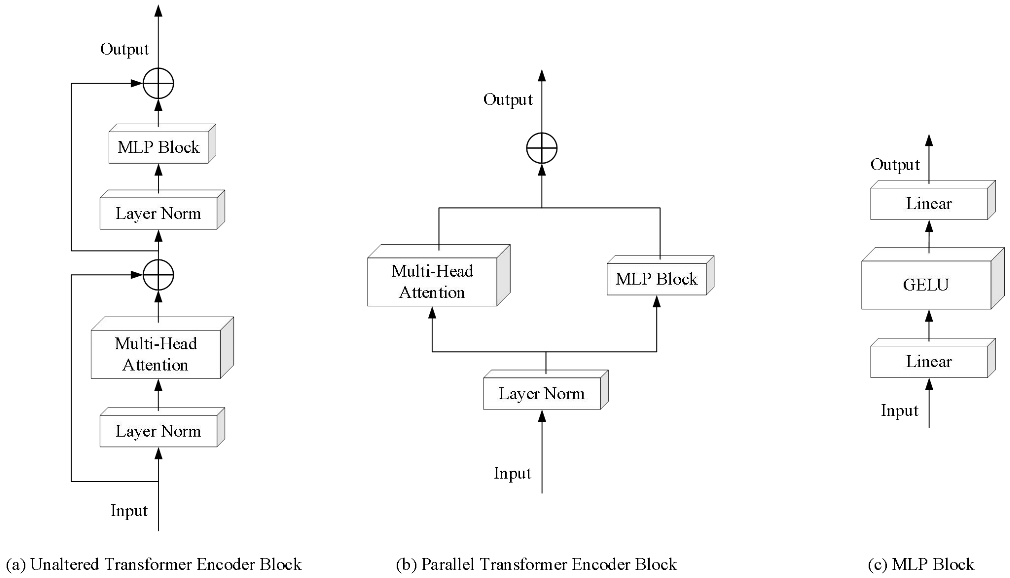


Figure 7. The structure diagram of PVTENet.

We introduce parallelization improvements at the Encoder Block level of the vision transformer, as shown in Figure 8. The structure of the MLP Block is shown in Figure 8. By improving the sequential structure of multi-head attention and the MLP Block to a parallel structure, we significantly improve computational efficiency and model throughput, enabling more efficient training of existing models.



(a) Unaltered Transformer Encoder Block

(b) Parallel Transformer Encoder Block

(c) MLP Block

Figure 8. The structure diagram of the Encoder Block. (a) is the unchanged structure diagram, (b) is the structure diagram after parallelization, and (c) is the structure diagram of the MLP Block.

4.3. Design of the convolutional decoder

To ensure compatibility with RENet and PVTENet, we design a convolutional decoder as the decoding component of E^2-ViTUNet (Figure 9), featuring symmetric architecture relative to RENet. Each upsampling stage employs a 2 × *bilinearupsampling* followed by two convolutional layers (3 × 3 kernels) and one 1 × 1 shortcut branch. Before upsampling, skip connections fuse encoder features from corresponding levels. Crucially, residual connections propagate information across stages, mitigating gradient vanishing/explosion while enhancing feature richness and detail fidelity in output images.

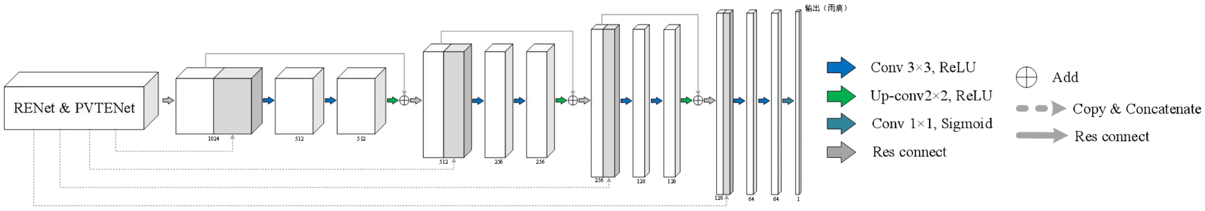


Figure 9. The structure diagram of the convolutional decoder.

4.4. *Composite loss function E^2-ViTUNet*

Currently, the mean squared error (MSE) is predominantly adopted as the loss function to quantify differences between rain-removed images and ground-truth references. This induces compromised detail reproduction—evident through attenuated texture gradients—which concurrently degrades deraining performance and scene authenticity. To mitigate these limitations, we formulate a composite loss function by integrating MSE with structural similarity (SSIM) [27], thereby balancing rain removal and background structural preservation.

We are given an input image I_i with rain, the output image $G(I_i)$ without rain, and the ground truth J_i. Therefore, a pixel-level MSE loss can be defined as:

$$L_{MSE} = \frac{1}{HWC} \sum_{x=1}^{H} \sum_{y=1}^{W} \sum_{z=1}^{C} \|G\left(I_i\right) - J_i\|^2 \tag{4.1}$$

where $H \times W \times C$ defines the spatial-channel tensor dimensions.

As a perceptual similarity metric, SSIM admits the formulation:

$$SSIM(G(I), J) = \frac{2\mu_{G(I)}\mu_J + C_1}{\mu_{G(I)}^2 + \mu_J^2 + C_1} \cdot \frac{2\sigma_{G(I)}\sigma_J + C_2}{\sigma_{G(I)}^2 + \sigma_J^2 + C_2} \tag{4.2}$$

where μ_x and σ_x^2 denote luminance statistics of image x, σ_{xy} is the cross-image covariance, and C_1, C_2 are regularization factors.

SSIM ∈ [0, 1] monotonically quantifies structural authenticity to ground truth, yielding the loss as:

$$L_{SSIM} = 1 - SSIM(G(I), J) \tag{4.3}$$

The compound loss $L_{E2-ViTUNet}$ integrates mean squared error with structural similarity metrics:

$$L_{E2-ViTUNet} = L_{MSE} + \lambda L_{SSIM} \tag{4.4}$$

where λ arbitrates the trade-off between photometric accuracy (L_{MSE}) and perceptual integrity (L_{SSIM}). Appropriately calibrated, this hybrid objective concurrently enforces pixel-wise fidelity and structural coherence, propelling the deraining network toward photorealistic output generation.

5. Preformance evaluation

This section details experimental settings and evaluates the proposed E²-ViTUnet.

5.1. Simulation settings

5.1.1. Datasets

Rain100H/L datasets. These benchmark datasets simulate precipitation gradients ranging from light rain to heavy rain. They provide rain-damaged input images and original reference images for training and evaluating image rain removal algorithms. Rain100H [28] contains 1800 training pairs and 200 test pairs, simulating dense and complex rainfall patterns such as rain streaks and mist, which induce severe visual occlusion and increased rain removal difficulty. Rain100L [28] includes 200 training pairs and 200 test pairs, characterized by sparse rain streaks with reduced visual interference, resulting in simpler rain removal compared to Rain100H. With standardized evaluation pairs and distinct difficulty levels, these datasets serve as essential benchmarks for assessing the robustness and generalization capabilities of image rain removal models.

Real-world rainy images dataset. To quantify real-world performance, we constructed a diversity-constrained benchmark comprising 300 publicly sourced rainy images, exhibiting heterogeneous scene semantics, precipitation densities, and directional rainfall patterns. Sample images are illustrated in Figure 10. This dataset serves exclusively as a test benchmark.

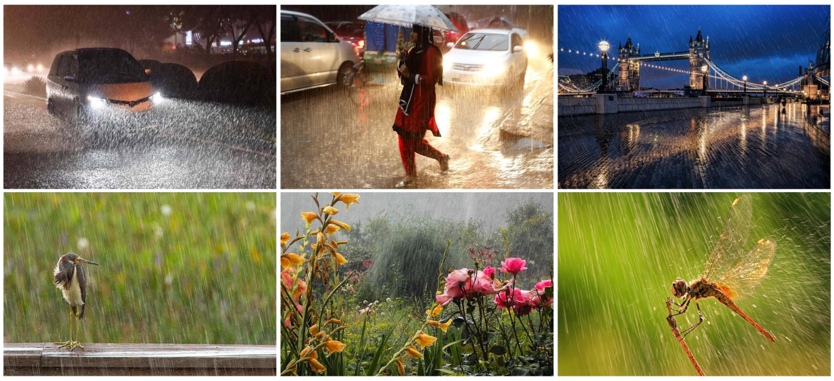


Figure 10. Sample images from the real-world rainy dataset.

5.1.2. Details and parameters of E^2-ViTUNet

The architecture was implemented in PyTorch on dual NVIDIA RTX 3090 GPUs, employing Adam optimization with batch size 8. Training protocol featured cosine-annealed learning rate decay: initialized at 10^{-3} for 3000 epochs with periodic halving at 25-epoch intervals. The SSIM regularization weight was tuned empirically to $\lambda = 0.2$, with identical computational configurations maintained throughout all experimental trials.

5.1.3. Quality measures

The comparative assessment framework integrates three quantitative fidelity metrics:

1) Peak Signal-to-Noise Ratio (PSNR) [29]

2) Structural Similarity Index (SSIM) [27]

3) Feature Similarity Index (FSIM) [30]

In absence of reference images for real data, multi-observer perceptual assessment was conducted for comparative analysis.

5.2. Experimental results

5.2.1. Evaluation on a synthetic dataset

With ground-truth references available for this evaluation suite, method performance was quantified using established perceptual metrics–Peak Signal-to-Noise Ratio, Structural Similarity Index, and Feature Similarity Index. Table 1 presents the comparative results based on these metrics, indicating that our proposed E^2-ViTUNet method significantly outperforms the others across PSNR, SSIM, and FSIM. Cumulatively, the framework demonstrates significant advancements beyond existing state-of-the-art deraining implementations.

Table 1. PSNR/SSIM/FSIM over Rain100H and Rain100L.

Datasets	Year	Rain100H			Rain100L		
Metrics		PSNR	SSIM	FSIM	PSNR	SSIM	FSIM
DSC [4]	2015	15.19	0.384	0.617	28.66	0.865	0.887
GMM [5]	2016	14.44	0.392	0.620	27.76	0.867	0.870
DDN [18]	2017	22.85	0.725	0.818	32.38	0.926	0.938
RESCAN [19]	2018	28.86	0.864	0.882	36.87	0.975	0.975
DualCNN [31]	2018	14.23	0.468	0.703	26.87	0.860	0.881
DID-MDN [32]	2018	22.97	0.732	0.821	32.44	0.931	0.942
PReNet [20]	2019	29.45	0.898	0.911	36.98	0.978	0.980
SPANet [33]	2019	25.11	0.833	0.856	35.33	0.969	0.953
DiG-CoM [12]	2020	22.40	0.705	0.805	33.43	0.937	0.944
RCDNet [21]	2020	31.26	0.908	0.925	38.82	0.984	0.986
MPRNet [13]	2021	30.41	0.890	0.889	36.40	0.965	0.921
SPDNet [34]	2021	32.73	0.925	0.939	41.78	0.988	0.989
EfDeRain [35]	2021	31.27	0.908	0.925	38.83	0.984	0.986
ESTINet [11]	2022	28.64	0.887	0.913	36.81	0.978	0.979
Restormer [36]	2022	31.46	0.904	0.914	38.99	0.978	0.910
MMRDNet [9]	2023	32.10	0.957	0.952	41.80	0.995	0.994
AUNet [8]	2024	29.80	0.899	0.912	38.70	0.968	0.951
DMRONet [10]	2024	30.81	0.949	0.945	41.18	0.994	0.993
Ours		**35.91**	**0.963**	**0.981**	**47.70**	**0.997**	**0.999**

Visual exemplars are provided to qualitatively contrast the proposed framework against contemporary state-of-the-art methods. Figures 11 and 12 demonstrate comparative deraining performance on the Rain100L and Rain100H benchmark datasets. Notably, these cases were strategically sampled from challenging scenarios to validate E^2-ViTUNet's robustness under complex precipitation conditions. First, in comparison with the prior-based method, the DSC [4] method performs inadequately, while E^2-ViTUNet can effectively remove rain streaks. Second, we also provide visual comparison results with three other classic deep learning methods. It is evident that the rain removal results of these three methods show residual precipitation patterns and processing artifacts, while compromising the fidelity of the background structure. Finally, visual comparison results with the latest method, DMRONet, are presented. In the comparison of the second and third samples, E^2-ViTUNet is observed to preserve finer iceberg textures and plant details. In contrast, our proposed method removes most rain streaks while preserving rich image details, yielding the cleanest results.

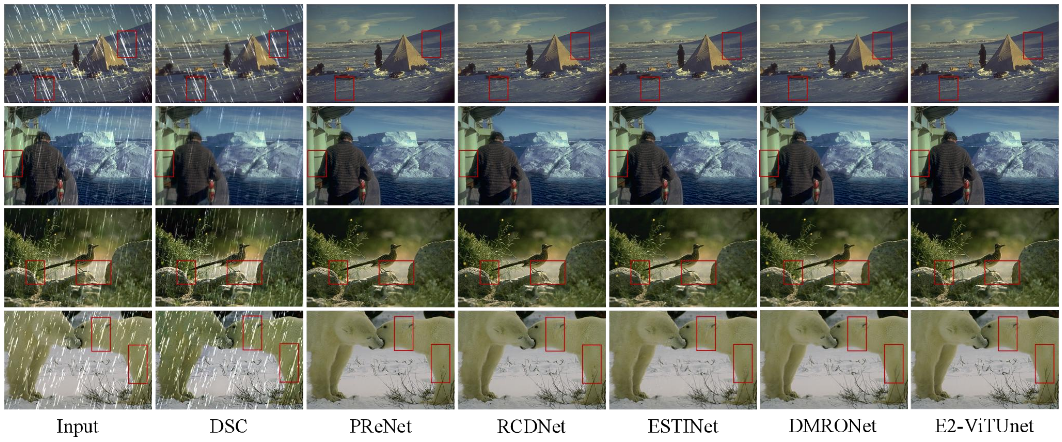


Figure 11. E^2-ViTUNet precipitation artifact clearance on Rain100L.

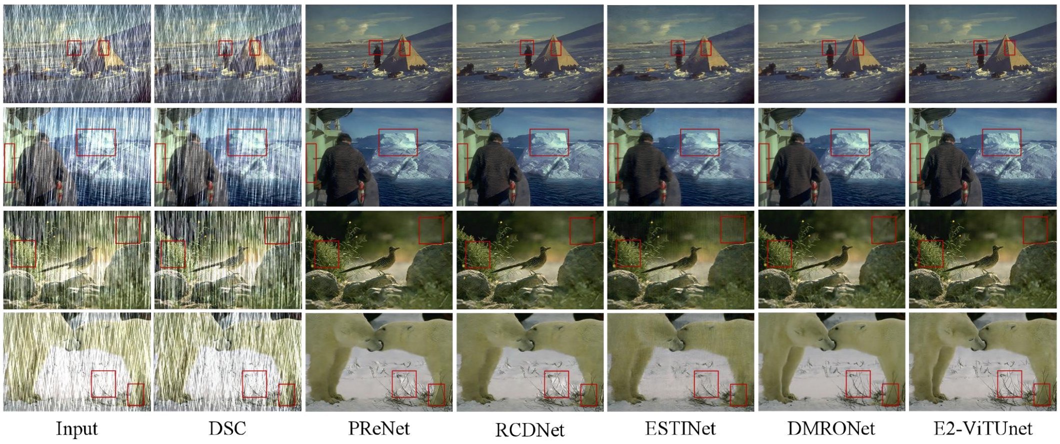


Figure 12. E^2-ViTUNet precipitation artifact clearance on Rain100H.

5.2.2. Evaluation on a real-world dataset

We collected various real rain scene images in different environments for evaluation to verify the effectiveness of E^2-ViTUNet. As shown in Figure 13, specific regions of interest (ROIs) are selected for visual comparison. The DSC method introduces artifacts during rain streak removal. Despite achieving reasonable visual quality, PReNet, RCDNet, ESTINet, and DMRONet exhibit residual rain streaks in the selected ROIs. In contrast, our method effectively removes most rain streaks while preserving background image details. The intuitive evidence in Figure 13 confirms the effective suppression of rain streaks by E^2-ViTUNet in real-world scenarios. Notwithstanding this limitation, our proposed approach delivers superior results relative to existing methods, yielding cleaner images with enhanced visual clarity.

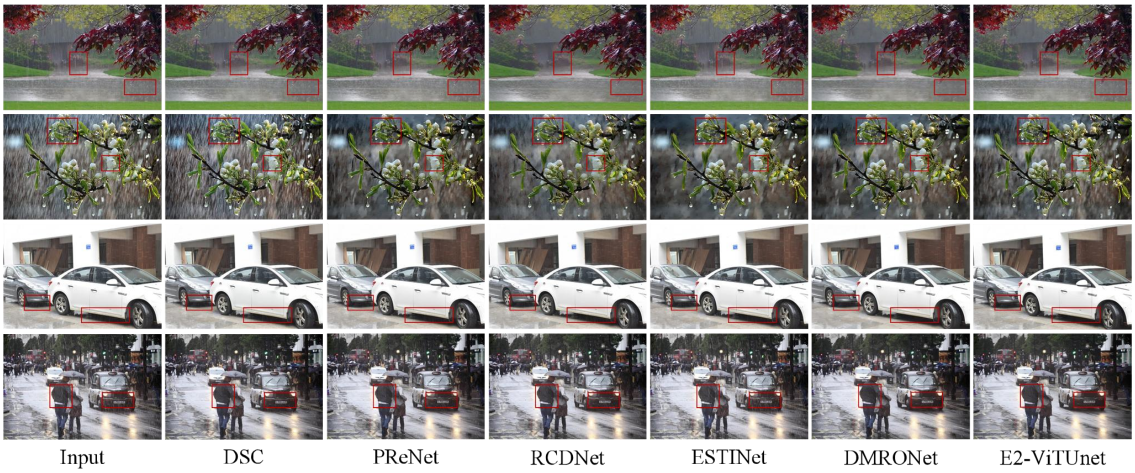


Figure 13. E^2-ViTUNet precipitation artifact clearance on the real-world rainy images dataset.

5.2.3. Evaluation on runtime and parameter numbers

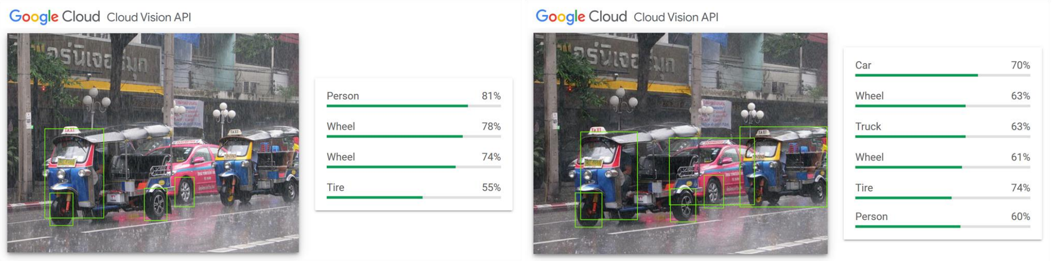
The computational efficiency of our framework was benchmarked against contemporary state-of-the-art methods. Quantitative profiling in Table 2 reveals significantly higher time complexity for DSC [4] and GMM [5]—a direct consequence of their iterative optimization mechanisms as generative modeling frameworks. PReNet [20] and ESTINet [11] exhibit lower computational costs, yet fail to effectively eliminate rain streaks. Regarding parameter count, while our algorithm is suboptimal, it remains highly competitive. Overall, our method strikes a superior balance between performance and computational efficiency.

5.2.4. Evaluation on object detection

We conducted experiments using the Google Cloud Vision API (https://cloud.google.com/vision/), and the results showed that E^2-ViTUNet can effectively mitigate the impact of rainy weather on outdoor visual systems while evaluating its practicality in computer vision applications. Figure 14 demonstrates clear resolution of background objects, such as the pink vehicle, in processed outputs. This indicates that images restored by E^2-ViTUNet not only enhance object detection metrics but also increase identifiable targets. Figure 15 documents API error correction: A flower stem mislabeled as an insect by Google Vision API was resolved through E^2-ViTUNet processing. The results reveal that restored images following rain-streak suppression achieve superior visibility restoration and significantly enhance object recognition performance in vision systems.

Table 2. Computational efficiency analysis of E^2-ViTUNet.

Methods	Year	Parameter (M×10^{-2})	Inference time (s)
DSC [4]	2015	–	99.2
GMM [5]	2016	–	371.2
DDN [18]	2017	0.06	0.6
RESCAN [19]	2018	0.05	0.6
DualCNN [31]	2018	0.69	20.2
PReNet [20]	2019	0.17	0.2
SPANet [33]	2019	0.28	0.4
DiG-CoM [12]	2020	4.10	2.8
RCDNet [21]	2020	3.17	2.7
MPRNet [13]	2021	3.64	1.0
SPDNet [34]	2021	3.04	0.4
ESTINet [11]	2022	0.43	0.3
MMRDNet [9]	2023	0.81	0.6
AUNet [8]	2024	–	0.3
DMRONet [10]	2024	0.25	0.8
Ours		0.21	0.9



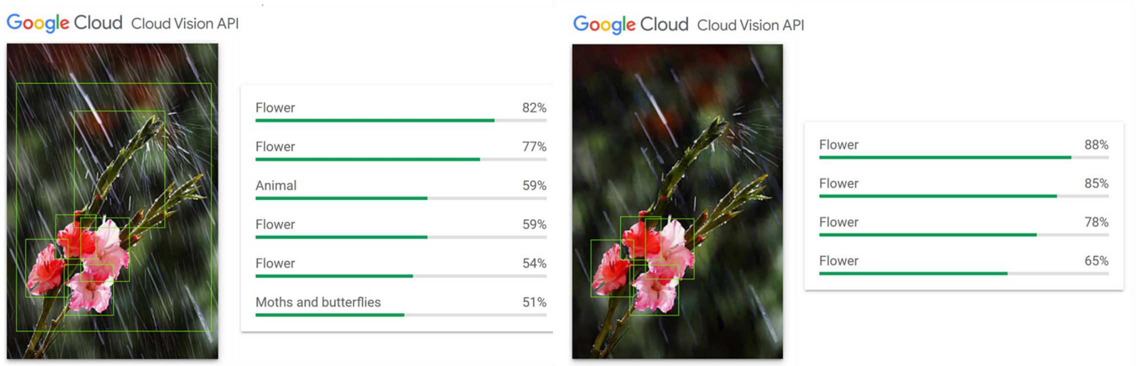
Person 81%
Wheel 78%
Wheel 74%
Tire 55%

Car 70%
Wheel 63%
Truck 63%
Wheel 61%
Tire 74%
Person 60%

(a) Recognizing result of original image

(b) Recognizing result of our removal result

Figure 14. Object detection analysis of E^2-ViTUNet. E^2-ViTUNet enhances primary object detection accuracy and recognition density.



Flower 82%
Flower 77%
Animal 59%
Flower 59%
Flower 54%
Moths and butterflies 51%

Flower 88%
Flower 85%
Flower 78%
Flower 65%

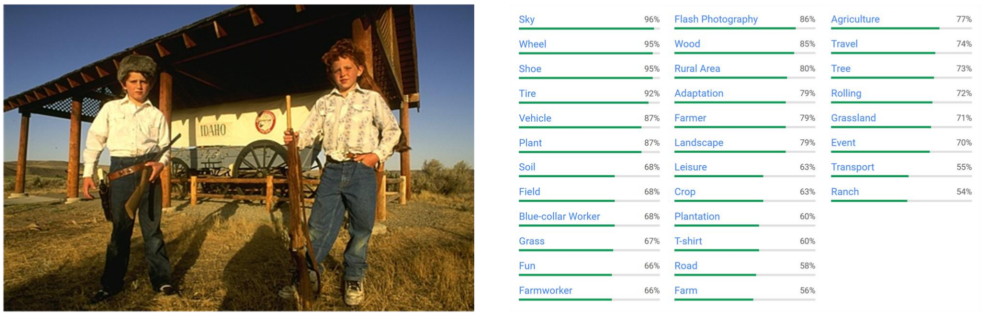
(a) Recognizing result of original image

(b) Recognizing result of our removal result

Figure 15. Object detection analysis of E^2-ViTUNet. E^2-ViTUNet corrects CV misclassification errors.

5.2.5. Evaluation on label detection

In addition to target recognition evaluation, we further conducted label detection evaluation on real rainfall images and rain-removed images generated by E^2-ViTUNet. Figure 16 delineates label-detection metrics, pairing object annotations with recognition confidence from Google Vision API processing.



Sky	96%	Flash Photography	86%	Agriculture	77%
Wheel	95%	Wood	85%	Travel	74%
Shoe	95%	Rural Area	80%	Tree	73%
Tire	92%	Adaptation	79%	Rolling	72%
Vehicle	87%	Farmer	79%	Grassland	71%
Plant	87%	Landscape	79%	Event	70%
Soil	68%	Leisure	63%	Transport	55%
Field	68%	Crop	63%	Ranch	54%
Blue-collar Worker	68%	Plantation	60%		
Grass	67%	T-shirt	60%		
Fun	66%	Road	58%		
Farmworker	66%	Farm	56%		

(a) A sample of label detection

(b) A result of label detection

Figure 16. Schematic of annotation-confidence pairing for label detection.

Ten original rainy images were randomly selected as samples. Quantitative comparison of detectable labels between original and restored images reveals that labels in E^2-ViTUNet restored images exhibit increased counts (Figure 17). The results demonstrate significant visibility improvement and enhanced label detectability after processing.

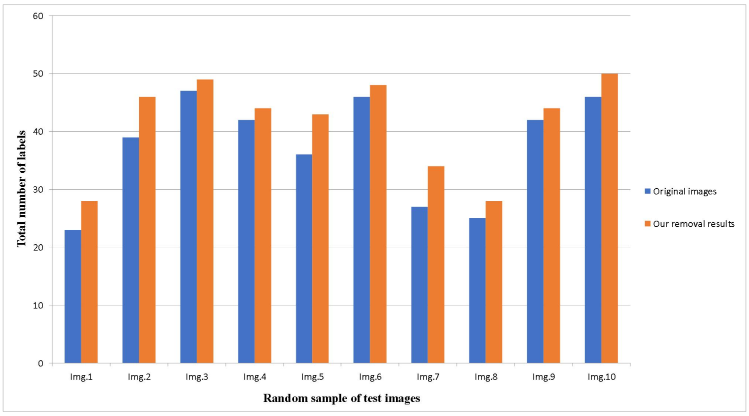
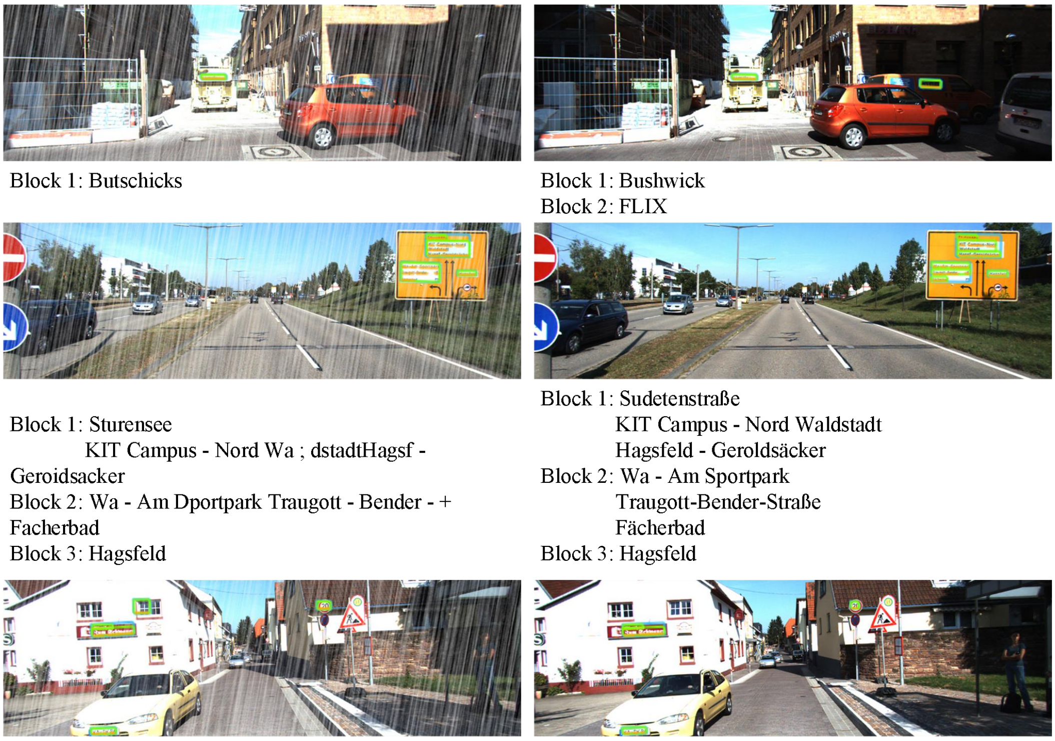


Figure 17. Lable detection analysis of E^2-ViTUNet. The evaluation corpus comprises ten real-world precipitation scenes, randomly curated for statistical validation.

5.2.6. Evaluation on text detection

Beyond label recognition evaluation, we performed detection and recognition of textual elements in images. As shown in Figure 18, our OCR analysis using the Google Vision API followed the experimental content established by Zheng et al [10]. Three representative cases were selected to illustrate experimental findings. The results demonstrate that E^2-ViTUNet-processed imagery enhances OCR precision for text extraction in the first sample. In the second sample, OCR successfully segments textual elements, indicating improved text discernibility through our algorithm.

The final sample reveals rain streak interference causes OCR to misidentify a window as H and a traffic signal's 20 as (20), whereas E^2-ViTUNet-processed images eliminate these errors. Empirical verification shows that compared to the experimental results of Zheng et al. [10], E^2-ViTUNet produces reality-confirmed outputs, as evidenced by verifiable geographic entities in real-world databases. Thus, our algorithm significantly enhances text legibility in degraded scenes.



Block 1: Butschicks

Block 1: Bushwick
Block 2: FLIX

Block 1: Sturensee
KIT Campus - Nord Wa ; dstadtHagsf -
Geroidsacker
Block 2: Wa - Am Dportpark Traugott - Bender - +
Facherbad
Block 3: Hagsfeld

Block 1: Sudetenstraße
KIT Campus - Nord Waldstadt
Hagsfeld - Geroldsäcker
Block 2: Wa - Am Sportpark
Traugott-Bender-Straße
Fächerbad
Block 3: Hagsfeld

Block 1: Wirtsband
Zum Schwanen
Block 2: KA - DK86
Block 3: H
Block 4: (20)

Block 1: Wirtshaus zum Schwanen
Block 2: KA - DK86
Block 3: 20

Figure 18. OCR analysis of E^2-ViTUNet. E^2-ViTUNet mitigates three persistent failure modes in scene text recognition.

5.3. *Ablation studies*

Ablation studies were performed to validate architectural improvements of the proposed method. For experimental efficiency, evaluations utilized the Rain100L and Rain100H datasets. Component-wise ablation studies were implemented wherein the complete architecture was decomposed into constituent modules, with systematic substitution of architectural components. As presented in Table 3, results from ablating encoder components demonstrate that integrating RENet and PVTENet yields substantial gains across quantitative metrics.

Concurrently, we conducted ablation studies on residual connections and CBAM in the RENet. PVTENet remained active during these experiments, as previous experiments have validated PVTENet's effectiveness, eliminating the need for PVTENet deactivation. Table 4 presents the ablation results for residual connections and CBAM. The data indicate that CBAM enhances model performance, attributed to improved feature representation of rain patterns. Notably, residual connections significantly improve model accuracy. We attribute this improvement to their ability to mitigate network degradation in deep architectures and facilitate gradient flow during training. Therefore, both residual connections and CBAM are essential components of E^2-ViTUNet.

Table 3. Network module ablation analysis of E^2-ViTUNet.

Networks		Rain100H			Rain100L		
RENet	PVTENet	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM
Yes	No	30.63	0.908	0.915	38.87	0.988	0.976
No	Yes	29.31	0.892	0.903	37.76	0.985	0.972
Yes	Yes	**35.91**	**0.963**	**0.981**	**47.70**	**0.997**	**0.999**

Table 4. Network module ablation analysis of E^2-ViTUNet.

RENet		PVTENet	Rain100H			Rain100L		
Residual connections	CBAM		PSNR	SSIM	FSIM	PSNR	SSIM	FSIM
No	No	Yes	32.13	0.958	0.952	42.00	0.994	0.994
No	Yes		33.28	0.959	0.961	44.92	0.995	0.994
Yes	No		35.05	0.961	0.978	47.16	0.996	0.996
Yes	Yes		**35.91**	**0.963**	**0.981**	**47.70**	**0.997**	**0.999**

6. Discussion

This section addresses applicability considerations as detailed below.

1) **Model stability.** Compared to traditional stacking-based approaches using extremely deep networks, the proposed E^2-ViTUNet mitigates gradient vanishing/explosion issues. RENet captures low-frequency background details through local receptive fields, while PVTENet processes global dependencies via self-attention. These pathways operate synergistically rather than through unstructured deepening of single-branch pathways. This heterogeneous dual-encoding architecture attenuates reliance on ultra-deep single architectures and suppresses gradient instabilities during training. Residual connections further ensure effective propagation of rain streak information. Consequently, the model preserves degraded background details while maintaining training stability, producing restored images with enhanced visual fidelity.

2) **Performance and robustness.** E^2-ViTUNet demonstrates superior performance in single-image rain removal, significantly enhancing image visibility and quality. Comprehensive evaluations across multiple datasets confirm its capability in processing varied rain streak densities, exhibiting considerable robustness. Both qualitative inspection and quantitative analysis validate its effectiveness for rain removal tasks. Critically, its capacity for improving image clarity has been validated in downstream computer vision applications, demonstrating deployment potential in security surveillance and autonomous driving domains while enabling real-time processing capabilities.

3) **Limitations.** While E^2-ViTUNet achieves compelling performance in handling rain streaks for single-image deraining, we observe that its effectiveness diminishes when addressing raindrop removal tasks. This limitation may stem from our substantial innovations being predominantly

concentrated in the encoder architecture, without commensurate enhancements to the decoder module. Consequently, the decoder's image reconstruction capability fails to match the encoder's sophisticated feature representation capacity. This represents an unresolved challenge in the current study, which we will systematically investigate in future work.

7. Industrial applications

Rainfall constitutes a prevalent phenomenon that compromises visibility and impacts vision systems in autonomous driving, security surveillance, and drone applications. Traditional deep rain removal models often suffer from critical background detail degradation due to training instabilities. The proposed E^2-ViTUNet effectively mitigates this core limitation: Its heterogeneous dual-encoder architecture integrates RENet's local representation learning with PVTENet's global modeling, while residual connections ensure effective propagation of rain streak information. This design effectively mitigates gradient instabilities during training, guaranteeing high-fidelity image restoration. The enhanced visual clarity provided by our model has been empirically validated in diverse vision applications, demonstrating significant potential for real-time deployment.

8. Conclusions

This work introduces an E^2-ViTUNet targeting single-image rain removal. The architecture incorporates a heterogeneous dual-encoding framework merging RENet's convolutional local feature extraction with PVTENet's attention-based global modeling. Employing residual pathways enables supplementary learning of high-frequency details, ensuring spatially coherent restoration outputs. Trained on synthetic datasets, the model demonstrates robust generalization capabilities on real-world rainy images, generating visually clear restored images. Quantitative validation using the Google Vision API quantitatively validates the efficacy for vision applications. Even under challenging rain conditions, the algorithm eliminates most rain streaks while preserving intricate background details. These results collectively demonstrate that E^2-ViTUNet achieves state-of-the-art rain removal performance with exceptional robustness, significantly advancing image clarity in adverse conditions.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Mid-Career Faculty Educational Research of Fujian under Grant JAT210853, the Vocational Education Research Project of Fujian under Grant ZJGA2024021, the Natural Science Foundation of Fujian under Grant 2023J01924, and the National Natural Science Foundation of China under Grant 62001199.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. H. Que, J. Weng, Y. Fang, K. Hu, H. Wei, Y. Xu, DCD-Net: Image deraining with delta convolution and joint calibration attention, *Signal Image Video Process.*, **19** (2025), 42. https://doi.org/10.1007/s11760-024-03682-4

2. M. Chen, P. Wang, D. Shang, P. Wang, Cycle-attention-derain: Unsupervised rain removal with CycleGAN, *Visual Comput.*, **39** (2023), 3727–3739. https://doi.org/10.1007/s00371-023-02947-2

3. S. Sun, S. Fan, Y. F. Wang, Exploiting image structural similarity for single image rain removal, in *2014 IEEE International Conference on Image Processing (ICIP)*, (2014), 4482–4486. https://doi.org/10.1109/ICIP.2014.7025909

4. Y. Luo, Y. Xu, H. Ji, Removing rain from a single image via discriminative sparse coding, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 3397–3405. https://doi.org/10.1109/ICCV.2015.388

5. Y. Li, R. T. Tan, X. Guo, J. Lu, M. S. Brown, Rain streak removal using layer priors, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2736–2744. https://doi.org/10.1109/CVPR.2016.299

6. Y. Chen, C. T. Hsu, A generalized low-rank appearance model for spatio-temporally correlated rain streaks, in *2013 IEEE International Conference on Computer Vision*, (2013), 1968–1975. https://doi.org/10.1109/ICCV.2013.247

7. L. Kang, C. Lin, Y. Fu, Automatic single-image-based rain streaks removal via image decomposition, in *IEEE Trans. Image Process.*, **21** (2012), 1742–1755. https://doi.org/10.1109/TIP.2011.2179057

8. Z. Zheng, Z. Chen, S. Wang, W. Wang, Dual-attention U-Net and multi-convolution network for single-image rain removal, *Visual Comput.*, **40** (2024), 7637–7649. https://doi.org/10.1007/s00371-023-03198-x

9. Z. Zheng, Z. Chen, S. Wang, W. Wang, H. Wang, Memory-efficient multi-scale residual dense network for single image rain removal, *Comput. Vision Image Understanding*, **235** (2023), 103766. https://doi.org/10.1016/j.cviu.2023.103766

10. Z. Zheng, Z. Chen, W. Wang, M. Huang, H. Wang, Dual parallel multi-scale residual overlay network for single-image rain removal, *Signal Image Video Process.*, **18** (2024), 2413–2428. https://doi.org/10.1007/s11760-023-02917-0

11. K. Zhang, D. Li, W. Luo, W. Ren, W. Liu, Enhanced spatio-temporal interaction learning for video deraining: Faster and better, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 1287–1293. https://doi.org/10.1109/TPAMI.2022.3148707

12. W. Ran, Y. Yang, H. Lu, Single image rain removal boosting via directional gradient, in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, (2020), 1–6. https://doi.org/10.1109/ICME46284.2020.9102800

13. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. Yang, et al., Multi-stage progressive image restoration, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 14816–14826. https://doi.org/10.1109/CVPR46437.2021.01458

14. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16×16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.

15. I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, **2** (2014), 3104–3112.

16. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

17. R. K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks, in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, (2015).

18. X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, J. Paisley, Removing rain from single images via a deep detail network, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1715–1723. https://doi.org/10.1109/CVPR.2017.186

19. X. Li, J. Wu, Z. Lin, H. Liu, H. Zha, Recurrent squeeze-and-excitation context aggregation net for single image deraining, in *Computer Vision-ECCV 2018*, **11211** (2018), 262–277. https://doi.org/10.1007/978-3-030-01234-2_16

20. D. Ren, W. Zuo, Q. Hu, P. Zhu, D. Meng, Progressive image deraining networks: A better and simpler baseline, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3932–3941. https://doi.org/10.1109/CVPR.2019.00406

21. H. Wang, Q. Xie, Q. Zhao, D. Meng, A model-driven deep neural network for single image rain removal, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 3100–3109. https://doi.org/10.1109/CVPR42600.2020.00317

22. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in *Proceedings of the IEEE*, **86** (1998), 2278–2324. https://doi.org/10.1109/5.726791

23. J. L. Elman, Finding structure in time, *Cognit. Sci.*, **14** (1990), 179–211. https://doi.org/10.1207/s15516709cog1402_1

24. R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in *Proceedings of the 30th International Conference on International Conference on Machine Learning*, **28** (2013), 1310–1318.

25. S. Woo, J. Park, J. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in *Computer Vision-ECCV 2018*, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

26. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 6000–6010.

27. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.*, **13** (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

28. W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, S. Yan, Deep joint rain detection and removal from a single image, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1685–1694. https://doi.org/10.1109/CVPR.2017.183

29. Q. Huynh-Thu, M. Ghanbari, Scope of validity of psnr in image/video quality assessment, *Electron. Lett.*, **44** (2008), 800–801. https://doi.org/10.1049/el:20080522

30. L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, *IEEE Trans. Image Process.*, **20** (2011), 2378–2386. https://doi.org/10.1109/TIP.2011.2109730

31. J. Pan, S. Liu, D. Sun, J. Zhang, Y. Liu, J. Ren, et al., Learning dual convolutional neural networks for low-level vision, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3070–3079. https://doi.org/10.1109/CVPR.2018.00324

32. H. Zhang, V. M. Patel, Density-aware single image de-raining using a multi-stream dense network, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 695–704. https://doi.org/10.1109/CVPR.2018.00079

33. T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, R. W. H. Lau, Spatial attentive single-image deraining with a high quality real rain dataset, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 12262–12271. https://doi.org/10.1109/CVPR.2019.01255

34. Q. Yi, J. Li, Q. Dai, F. Fang, G. Zhang, T. Zeng, Structure-preserving deraining with residue channel prior guidance, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 4218–4227. https://doi.org/10.1109/ICCV48922.2021.00420

35. Q. Guo, J. Sun, F. Juefei-Xu, L. Ma, X. Xie, W. Feng, et al., EfficientDeRain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 1487–1495. https://doi.org/10.1609/aaai.v35i2.16239

36. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. Yang, Restormer: Efficient transformer for high-resolution image restoration, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 5718–5729. https://doi.org/10.1109/CVPR52688.2022.00564