IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 August 2023, accepted 29 August 2023, date of publication 1 September 2023, date of current version 12 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311138

RESEARCH ARTICLE

Degradation-Aware Transformer for Single Image Deraining

PEIJUN ZHAO AND TONGJUN WANG
School of Information Engineering, Xinyang Agriculture and Forestry University, Xinyang 464000, China

Corresponding author: Tongjun Wang (wangtj@xyafu.edu.cn)

This work was supported by the Science and Technology Department of Henan Province under Grant 222102110189 and Grant 212102210534.

ABSTRACT The crux of image deraining originates from recognizing various rain patterns. Most existing methods employ low-level spatial or sequential information to reconstruct the rain-free background image. However, due to the lack of sufficiently capturing long-term contextual relation between pixels, these methods often lead to incompletely modeling rain patterns and visible rain residues remained. In this paper, we propose a novel Degradation-Aware Transformer (DAT), which leverages a multi-level contrastive learning to obtain discriminative degradation representations by a degradation-aware model. Based on this, we also design a degradation-aware self-attention mechanism to improve the restoration performance on diverse rain patterns. Benefiting from the developed self-attention mechanism, DAT is able to capture long-term relations between pixels as well as completely modeling the rain degradation patterns. Extensive experiments demonstrate that our DAT is able to achieve the state-of-the-art performance of single image deraining in terms of both qualitatively visual comparison and quantitative comparison with other baseline methods on benchmark datasets.

INDEX TERMS Image deraining, degradation representation, contrastive learning, generative model.

I. INTRODUCTION

Image deraining is a crucial task in low-level vision while capturing images in real-life scenes. Rain often causes undesired artifacts, blur, and content loss. More importantly, the rainy condition will make the performance of vision systems, such as surveillance cameras or visual depth estimation, degrade significantly. The main challenge in the image deraining task stems from the difficulty of recognizing complex rain patterns in real-world scenes. Thus, in this paper, we aim to design a strategy for deep convolutional networks to model various rain patterns and propose a generalizable framework to remove rain streaks thoroughly.

Image deraining degradation is mathematically defined by an additive form that a rain-contaminated image I is linearly composed of a rain layer R and a background layer B: $I = \alpha B + \beta R$, where α and β are the coefficients. The image deraining task is a highly ill-posed problem because there are infinite solutions for a rain-contaminated image I. Under such physical model, traditional methods for image deraining assume prior knowledge referring to the characteristics of rain patterns, such as sparse prior [1], [2], or low-rank decomposition [3], [4], to constrain the solution space. Besides, some methods [5], [6] establish a dictionary of rain patterns according to their physical property in the rainy image, such as regional degradation or different rain densities. Such prior-based methods are able to restore clean background images from rainy images with simple rain patterns yet failing to model complex rain patterns and often leave visible rain residues in the restored images.

Recent single image deraining methods gradually achieve promising results due to the powerful capability of feature learning by convolutional neural networks(CNNs). The typical CNN-based methods [7], [8], [9] aim to design frameworks that follow an encoder-decoder manner and perform the fully supervised learning on the output of CNNs to implicitly learn and remove the rain patterns. However, such plain CNN-based methods often leave visible rain or additional artifacts due to the lack of explicit modeling of rain patterns. To relieve the modeling problem of rain patterns, some methods [10], [11], [12], [13] leverage the attention mechanism [14] to model rain patterns in the latent feature

The associate editor coordinating the review of this manuscript and approving it for publication was Syed Islam.

space. Even so, the performance of rain modeling still relies a lot on the per-pixel supervision between the output images of deep network and groundtruths, leading to little improvement of deraining performance. In order to remove rain constituents more thoroughly, some methods [15], [16], [17] attempt to learn degradation representations of rain patterns from implicit cues of predicted degraded rainy regions to improve the performance of image deraining. However, this type of deraining methods often leaves visible rain degradations, resulting from the lack of explicitly semantic understanding in the latent feature space. Furthermore, although some recent methods attempt to model degradation patterns using Transformers, their performance still remains unsatisfactory due to the lack of explicit guidance on distinguishable degradation patterns.

Contrastive learning has brought in prominent performance improvement on the mainstream computer vision tasks [18], [19], [20] recently for its excellent performance on clustering various semantics. Specifically, contrastive learning pushes similar content close yet dissimilar content away by establishing positive and negative pairs for judgment. Though it works well on many high-level vision tasks, it is still not fully exploited in low-level vision tasks such as image deraining mainly because of the difficulty of defining effective contrastive pairs to let networks accurately model complex rain patterns. Some researches [21], [22] have attempted to employ contrastive learning by fully copying the manner in high-level vision tasks, but the high-level modeling scheme is hard to benefit the process of image reconstruction. This is mainly because contrastive learning in image deraining requires multi-level content contrast in the latent feature space to provide more semantic cues for rain separation.

In this paper, we propose a Degradation-Aware Transformer (*DAT*), which is a self-attention based framework for single image rain streak removal. It obtains discriminative degradation-aware representations by performing a novel multi-level contrastive learning mechanism, which is able to learn diverse rain patterns across different latent feature spaces. Then, we leverage these representations to improve the performance of the deraining Transformer by a novel degradation-aware self-attention mechanism. The overall contribution of our method is as follows:

- We propose a novel Degradation-Aware Transformer (DAT), which leverages degradation representations to completely model diverse rain patterns. To this end, we design a degradation-aware self-attention mechanism to restore rainy images under the guidance by their corresponding degradation representations.
- We design a multi-level contrastive learning mechanism for degradation-aware model to learn a discriminative latent feature space, and thus the latent representations of images with similar rain patterns could be clustered unsupervisedly.
- Extensive experiments on four benchmark datasets for single image deraining demonstrate the superiority of our proposed *DAT* over present state-of-the-art methods in both visual and quantitative performance.

II. RELATED WORK

There are a substantial researches on image deraining, and we select the most related works for a comprehensive review. In this section, We mainly review three main parts: prior knowledge-guided methods for image deraining, CNN-based methods for image deraining, and contrastive learning.

A. PRIOR KNOWLEDGE-GUIDED METHODS FOR IMAGE DERAINING

Most traditional researches on image deraining assume some prior knowledge to optimize the rainy image according to the mathematical definition of the deraining problem in a non-learning manner. Considering the frequency separation, Kang et al. [23] separate the rainy image into high- and low-frequency components, and recognize rain by its structural characteristics. Luo et al. [2] separate the rain layer and background layer, referring to the mutual structural information by sparse coding. Deng et al. [24] employ the sparsity and direction of rain patterns as prior knowledge to restore the background image. Chen and Hsu [3] design a low-rank representation model to extract the relevant structure of rain patterns in rainy images. Kim et al. [25] propose to leverage non-local means to detect the rain regions by kernel regression. Wang et al. [26] use a hierarchy scheme to iteratively recognize the rain patterns in the high-frequency domain of rainy images. Li et al. [27] employ the gaussian mixture model to separate the rain layer and background layer. Such traditional methods for single image deraining are able to remove rain streaks that are with simple patterns and similar characteristics, but often fail to satisfy the variety of rain patterns in real-world rainy images.

B. CNN-BASED METHODS FOR IMAGE DERAINING

With the fast development of deep CNNs, the performance of CNN-based methods for computer vision tasks has made prominent progress benefiting from their powerful ability of feature learning. At the same time, the generative models such as Generative Adversarial Nets(GAN) [29] or Variation Auto-Encoder(VAE) [30] have significantly promoted the quality of generated images, and thus many CNN-based methods are proposed for image deraining. Fu et al. [8] design a deep detailed network to remove rain patterns in the high-frequency domain, and then finally reconstruct the background and rain layers. Yang et al. [31] leverage the recurrent network to model the variety of rain patterns, such as density or directions, and remove rain constituents. Inspired by the non-local means method, Li et al. [32] design a non-local block in CNN models to obtain long-term context and facilitate rain streak removal. To further consider the influence of rain density, Zhang and Patel [33] propose a density-aware image deraining method using a multi-stream dense network, which removes rain components and estimates the rain density simultaneously.

To learn degradation representations and thoroughly remove diverse rain patterns, some methods [15], [16], [17] try to provide representations of degradation properties

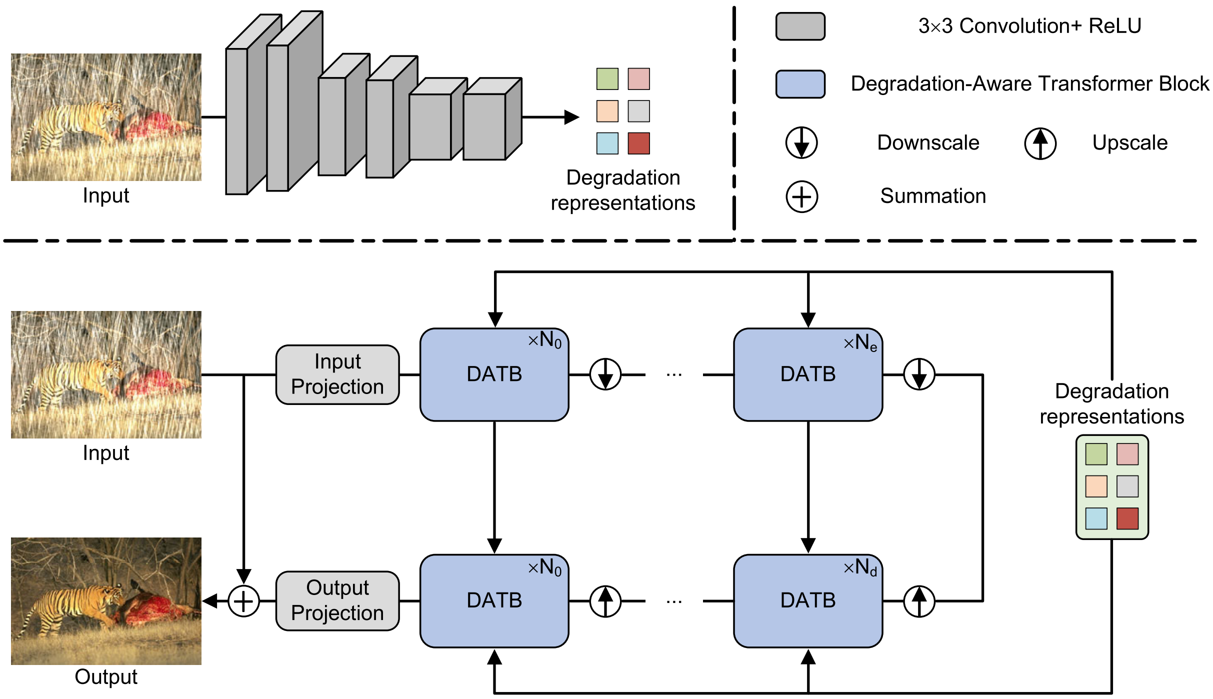


FIGURE 1. Given a rainy image, our *DAT* model first obtains the degradation representations of the rainy image as prior knowledge and then employs them to strengthen the restoration performance of deraining Transformer.

as implicit cues. For instance, Wang et al. [15] propose a context-enhanced representation learning and deraining network to effectively learn from predicted rainy regions. Nevertheless, as only limited degradation information is employed, such scheme often leads to insufficient modeling of rainy patterns. In summary, compared to prior knowledge-guided methods, CNN-based methods have achieved prominent improvements in deraining performance. However, this type of methods relies a lot on the fully supervised learning scheme and often appears to overfit the training dataset. Thus, it is in demand to explicitly model the difference between the background content and rain patterns.

C. CONTRASTIVE LEARNING

To improve the ability of feature representation by deep models, contrastive learning has been recently introduced into many computer vision tasks [19], [35] and achieves prominent performance. The main goal of contrastive learning is to push similar content close and keep dissimilar content far apart. Specifically, by recognizing patch pairs from different domains (augmented domain or unrelated domain) in a classification manner, deep models are able to cluster similar semantics in the latent feature space. Wang et al. [36] leverage contrastive learning to obtain representations of degradation for blind image super-resolution. Ji et al. [22] propose a transformer-based framework for image restoration, and introduce a multi-view contrastive learning mechanism. Wu et al. [21] employ the constrative learning mechanism to separate the haze component from the background image. However, most contrastive learning methods are designed for high-level vision tasks, such as image classification, detection, or segmentation, yet there are few works for the low-level vision tasks due to the difficulty of image reconstruction. Thus, we aim to propose a multi-level contrastive learning mechanism in this paper to effectively leverage the clustered degradation representations and leverage them to improve the performance of single image deraining. To this end, our model is based on Transformer to sufficiently model complex rain patterns and remove them thoroughly.

III. METHOD

In this section, we first overview the main framework of the proposed *DAT* for single image deraining. Then, we elaborate two specialized models, the designed degradation-aware model and the deraining transformer. Next, we introduce the detailed loss functions for effective supervision during the training process.

A. FRAMEWORK OVERVIEW

As shown in Figure 1, our *DAT* model includes two main parts, a degradation-aware model and a deraining transformer. The degradation-aware model aims to generate discriminative latent representations of the input rainy image, which serves as prior knowledge to further improve the performance of the deraining transformer.

To achieve this, it first employs stacked convolutional layers to map the input rainy image into the latent space, and performs representation learning on the obtained representations via contrastive learning. Thus, the representations can become more discriminative in an unsupervised manner. In this way, the learned representations are able to provide effective degradation information of the degraded image and guide the deraining Transformer to restore diverse rain patterns. Specifically, our *DAT* designs a Degradation-Aware Transformer Block (DATB) to absorb the prior knowledge of degrdation patterns from the degradation-aware model by the Degradation-Aaware Multi-head Self-Attention (DA-MSA) mechanism. Finally, our *DAT* is able to generate the realistic rain-free image from the degraded rainy image.

B. DEGRADATION-AWARE MODEL

As presented in Figure 1, the degradation-aware model is composed by stacked convolution layer and ReLU activation. Thus, the input rainy image **I** is able to be mapped into the latent feature space:

$$e = f_{\mathrm{DA}}(I), \tag{1}$$

where f_{DA} is the Degradation-Aware model, and **e** is the obtained representations.

The contrastive learning mechanism has benefited the representation learning of feature maps for many computer vision tasks [18], [19]. Unlike previous fully supervised learning, contrastive learning seeks to cluster similar semantics by high-level modeling in the latent feature space. In detail, the core strategy of typical contrastive learning establishes positive pairs by sampling image patches from the input image and its augmented image, and negative pairs by sampling image patches from other images. By reusing the encoder to classify the label of contrastive pairs, the deep model is able to obtain more effective semantic representations. Although it is able to enhance the performance of high-level vision tasks, the challenge of performing contrastive learning in low-level vision is still not fully exploited. It is mainly because of the requirement for sufficient information aggregation during image reconstruction.

Notably, our degradation-aware model performs contrastive learning on representations **e** in order to make the representations discriminative. To be specific, we propose the multi-level contrastive learning mechanism that performs contrastive learning on the representations of the output of each layer in degradation-aware model to comprehensively model rain patterns in various latent feature spaces. Given a rainy image I, the positive patches x_p are first sampled from both the restored image $\hat{I}$ of our degradation-aware model and groundtruth image I_{gt}. The negative patches x_n are sampled from both the input rainy image and restored image $\hat{I}$. By performing supervision on label classification, the deep model is able to know whether current semantics in input images contains rain patterns. In detail, we map the query patches, positive patches, and negative patches into vectorial query codes e_q, e_p, and e_n by the degradation-aware model. Typical contrastive learning fails to significantly enhance the performance of image restoration because of two main reasons. 1) high-level cues are difficult to explicitly guide the downstream image reconstruction tasks; 2) the reconstruction quality relies on the multi-level information in various feature spaces.

To solve above problems, we design a multi-level contrastive learning mechanism. As shown in Figure 1, the main consideration of our work is to perform contrastive learning in multi-level feature spaces and employ multi-level information extracted in the degradation-aware model to more completely discriminate rain patterns. The loss function of our multi-level contrastive learning mechanism is defined as follows:

$$\mathcal{L}_{\mathrm{mcl}} = -\sum_{l=1}^{n_l}\sum_{i=1}^{n_i}\log\left[\frac{\sum\limits_{j=1}^{n_j}\exp(\mathbf{e}_q^i\cdot\mathbf{e}_p^j)}{\sum\limits_{j=1}^{n_j}\exp(\mathbf{e}_q^i\cdot\mathbf{e}_p^j)+\sum\limits_{k=1}^{n_k}\exp(\mathbf{e}_q^i\cdot\mathbf{e}_n^k)}\right], \tag{2}$$

where n_i, n_j, and n_k denote the number of query patches, positive patches, and negative patches, respectively, and n_l is the l-th layer in the degradation-aware model. In this way, the representations of diverse rain patterns can be discriminative in the learned feature space, and thus the deraining Transformer can restore the target degradation by self-attention mechanism under the guidance of obtained representations.

C. DERAINING TRANSFORMER

The deraining Transformer follows an encoding-decoding framework to reconstruct the rain-free image from the rainy image. To be specific, the rainy image is first encoded into latent features by stacked Degradation-Aware Transformer Block (DATB). The DATB is based on the shifted window-based Transformer block [49], but has two main differences: 1) DATB employs degradation-aware multi-head self-attention mechanism to leverage the learned discriminative representations as prior knowledge while restoring rain patterns; 2) It employs depth-wise and point-wise convolutions to replace typical convolution layers.

Specifically, it employs stacked DATBs to progressively restore the rain patterns in the latent feature space under the guidance of degradation representations.

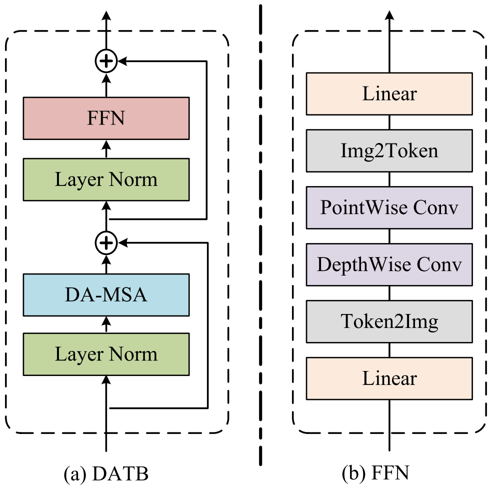
$$F_i = \sum_{i}^{N_e}\mathrm{DATB}_i(F_{i-1}, \mathrm{DR}), \tag{3}$$

where DR denotes the discriminative degradation representations, F$_i$ is the i-th DATB in the deraining Transformer, N_e is the number of DATBs.

1) DEGRADATION-AWARE TRANSFORMER BLOCK

While using Transformer for image restoration, there are two primary issues to overcome. Initially, the standard Transformer architecture [14] calculates self-attention across all tokens globally, resulting in a quadratic computational expense proportional to the number of tokens. This Characteristic is not suitable for high-resolution feature maps due to the large amount of pixels, making global self-attention expensive. Second, degradation information plays a crucial role in image restoration tasks since the degradation patterns are diverse and hard to be comprehensively modelled.

To relieve the aforementioned issues, we introduce a Degradation-Aware Transformer Block (DATB), as presented in Figure 2, which benefits from the self-attention mechanism to obtain contextual information, as well as the degradation guidance by the degradation-aware model to improve



(a) DATB

(b) FFN

FIGURE 2. The structure of degradation-aware transformer block.

the performance of the deraining Transformer. Specifically, given the input feature **X**, DATB is built based on the degrdation-aware multi-head self-attention mechanism and Feed-Forward Network (FFN):

$$\begin{aligned} X' &= \text{DA-MSA}(\text{LN}(X, \text{DR})) + X, \\ X_{\text{out}} &= \text{FFN}(\text{LN}(X')) + X', \end{aligned} \tag{4}$$

herein, LN is the layer normalization [51], DR is the learned representations by contrastive learning, DA-MSA is the degradation-aware multi-head self-attention mechanism, and X' denotes the intermediate feature maps. It is worth mentioning that we employ point-wise and depth-wise convolution layers in FFN to replace vanilla convolution inspired by the MobileNet [50], which enjoys similar performance to plain convolution but less computational overhead.

2) DEGRADATION-AWARE MULTI-HEAD SELF-ATTENTION

In order to significantly reduce the computational cost in image restoration tasks, we replace vanilla self-attention mechanism with the window-based self-attention [49]. Given the input features $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we partition **X** into non-overlapped windows of size M×M, and each window is further subsequently flatten and transposed. Then, self-attention is employed for feature of each window. As presented in Figure 3, the self-attention mechanism can be formulated as follows

$$\begin{aligned} X &= \{X^1, \ldots, X^N\}, N = HW/M^2, \\ Y &= \text{MSA}(\text{DR}W_k^{\text{DR}} + XW_k^Q, XW_k^K, XW_k^V) \\ \hat{X}_k &= \{Y_k^1, \ldots, Y_k^1\}, \end{aligned} \tag{5}$$

herein, **W** denotes the projection matrix for the k-th head. Next, the output features from all heads are concatenated and the process of multi-head attention can be formulated as follows

$$\text{MSA}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}} + B)V, \tag{6}$$

where d is the scaling factor and **B** is the position encoding.

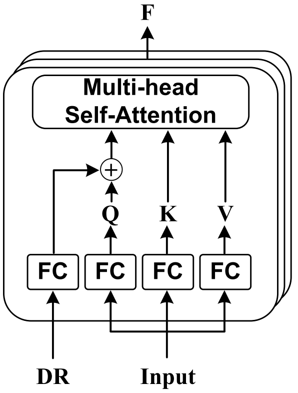


FIGURE 3. The structure of degradation-aware multi-head self-attention mechanism.

D. LOSS FUNCTIONS FOR PARAMETER OPTIMIZATION

We optimize the parameters of the proposed *DAT* by performing four types of loss functions as supervision signals. Besides the proposed multi-level contrastive learning loss $\mathcal{L}_{\text{mcl}}$, we perform other three loss functions, the reconstruction loss, the adversarial loss, and high-frequency loss, for model learning.

1) ADVERSARIAL LOSS

We employ the Markovian discriminator [38] to perform adversarial supervision and make the distribution of the restored images as realistic as the real-world background images. Generator G_{DAT} and Discriminator D are trained by solving *arg min*$_G$ *max*$_D$ $L_{\text{adv}}(G, D)$:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\text{I}\backsim\mathbb{P}_{\text{I}_{\text{gt}}}}[\log D(\text{I}_{\text{gt}})] + \mathbb{E}_{\text{I}\backsim\mathbb{P}_{\text{DAT}}}[1 - \log(D(G(\text{I})))], \tag{7}$$

where I and I_{gt} are the input rainy image and groundtruth background image.

2) HIGH-FREQUENCY LOSS

High-frequency information is essential to the reconstruction of the background image, and facilitates the modeling of rain patterns. Thus, we employ an edge detector [39] to obtain the high-frequency information of the restored image and groundtruth image, and perform supervision:

$$L_{\text{high}} = ||f(\text{I}_{\text{gt}}) - f(\hat{\text{I}})||_1, \tag{8}$$

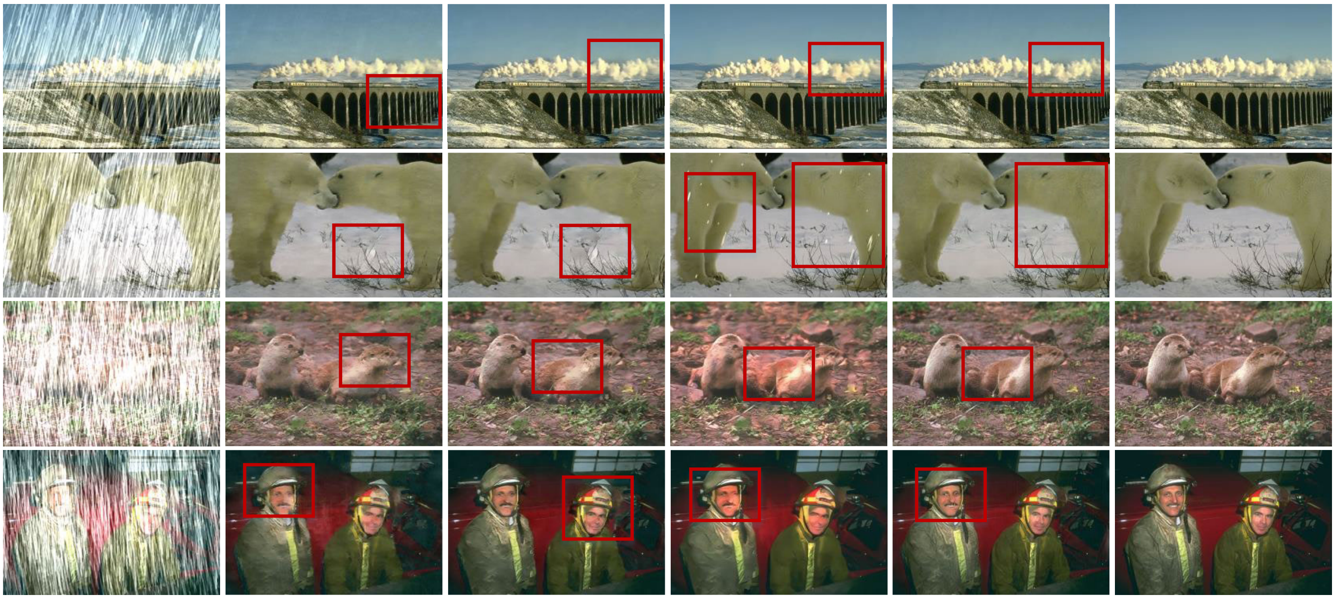
where f is the Canny detector which extracts the high-frequency information from RGB images.

3) RECONSTRUCTION LOSS

The reconstruction loss in our framework includes the pixel-wise reconstruction loss L_{pix} and the semantically perceptual loss L_{per}:

$$L_{\text{rec}} = L_{\text{pix}} + L_{\text{per}}. \tag{9}$$

The pixel loss measures the L_1 distance between the restored rain-free image and groundtruth on each pixel, which is



(a) Input (b) RESCAN (c) PReNet (d) MPRNet (e) DAT (ours) (f) Groundtruth

FIGURE 4. Visual comparison on test images for rain streak removal between our proposed *DAT* and competing state-of-the-art methods. Best viewed in zoom-in mode.

TABLE 1. Quantitative results of different models for rain streak removal in several synthetic datasets in terms of PSNR and SSIM.

Method	Rain100H [48]		Rain100L [48]		Rain800 [49]		Rain1400 [8]		Test1200 [35]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DDN [8]	18.27	0.524	29.12	0.908	26.64	0.848	27.87	0.890	27.52	0.882
DIDMDN [33]	16.84	0.512	27.39	0.802	22.45	0.730	25.01	0.745	24.69	0.737
CFDNet [46]	26.88	0.802	35.33	0.960	29.83	0.858	31.07	0.887	25.68	0.803
RESCAN [28]	28.95	0.884	37.68	0.970	30.95	0.876	32.41	0.892	31.45	0.884
PReNet [48]	27.64	0.865	36.99	0.963	29.56	0.860	31.23	0.917	32.02	0.899
RCDNet [45]	30.69	0.898	38.15	0.976	31.47	0.908	33.60	0.940	32.27	0.930
MSPFN [41]	29.43	0.887	37.63	0.970	31.52	0.912	32.48	0.923	32.53	0.932
DMGN [47]	30.72	0.903	38.53	0.983	31.91	0.917	34.06	0.939	30.46	0.892
MPRNet [43]	28.60	0.870	34.96	0.959	28.71	0.873	31.87	0.925	31.34	0.891
DAT(ours)	**31.09**	**0.903**	**39.54**	**0.984**	**32.49**	**0.919**	**34.39**	**0.946**	**33.65**	**0.938**

defined as follows:

$$L_{\text{pix}} = ||\text{I}_{\text{gt}} - \hat{\text{I}}||_1. \qquad (10)$$

The perceptual loss [40] learns consistency between restored image and groundtruth at the semantic level, which is defined as follows:

$$\mathcal{L}_{\text{per}} = \sum_{l=1}^{L} \|f^{\text{vgg}}(\text{I}_{gt}) - f^{\text{vgg}}(\hat{\text{I}})\|_1, \qquad (11)$$

where $f^{\text{vgg}}(\hat{\text{I}})$ and $f^{\text{vgg}}(\text{I}_{gt})$ are the feature maps extracted by the pre-trained VGG-19.

The overall loss functions are defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{high}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{adv}} + \lambda_4 \mathcal{L}_{\text{mcl}}, \qquad (12)$$

where λ_0, λ_1, λ_2, and λ_3 are hyper-parameters to balance between different losses. In our experiments, we empirically set $\lambda_1 = \lambda_4 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$.

IV. EXPERIMENTS

In this section, we design sufficient experiments to manifest the effectiveness of our proposed methods for single image deraining. First, we elaborate the implementation details in our experiments, including datasets, evaluation metrics, and experimental results. Then, we analyze the quantitative and qualitative results compared with the state-of-the-art methods for image deraining. Finally, we conduct the ablation study to investigate the effectiveness of each technical component in our *DAT*.

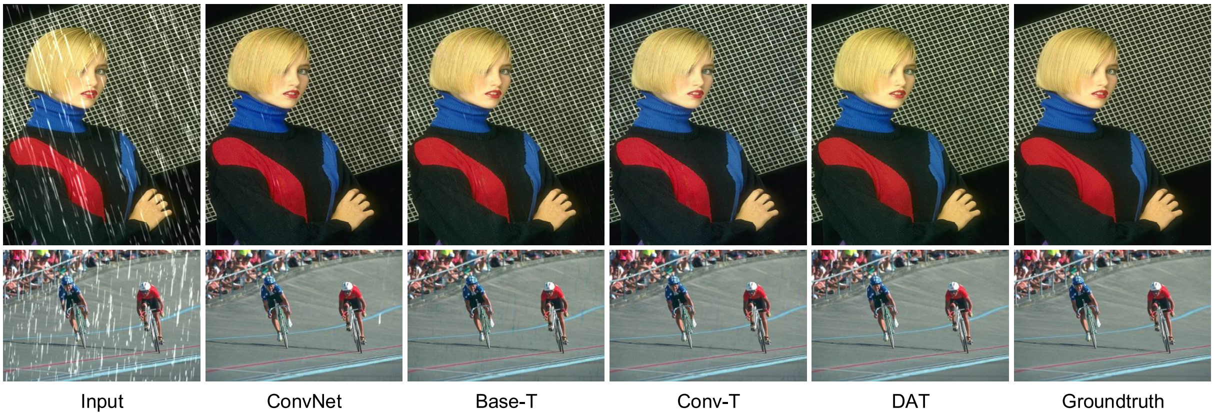


FIGURE 5. **Visual results of ablation study by four variants of the proposed DAT model. Best viewed in zoom-in mode.**

A. EXPERIMENTAL SETTINGS

1) DATASETS

We evaluate our proposed *DAT* on four benchmark datasets for single image deraining, including Rain100 [41], Rain800 [42], Rain1400 [8], and Test1200 [33]. Rain100 [41] includes two parts, Rain100L(100 light rain images) and Rain100H(100 heavy rain images), which consider the density of rain streaks. Rain800 [42] contains 700 training image pairs and 100 images for testing. Rain1400 [8] has 12600 image pairs for training and 1400 images for testing. Test1200 [33] has 1200 images for testing.

2) IMPLEMENTATION DETAILS

We conduct the experiments in this section by using 4 RTX 3090 GPUs under the Pytorch framework of deep learning. The optimizer for stochastic gradient descent is the Adam [44], the learning rate is 0.001, and altogether 100 epochs are used for model training.

3) EVALUATION METRICS

Two main metrics are employed to quantitatively evaluate the results of our *DAT* and competing state-of-the-art methods for image deraining, the Peak Signal-to-Noise Ratio (PSNR) and the Structural SIMilarity (SSIM) between the restored image and groundtruth image. Importantly, higher PSNR and SSIM denote the better quality of restored rain-free images.

B. EXPERIMENTAL RESULTS

In this section, we analyze the restoration performance by our *DAT* and other state-of-the-art methods for single image rain streak removal in terms of quantitative evaluation metrics and qualitative visual quality.

1) QUANTITATIVE COMPARISON

We compare the deraining results of our proposed method with other nine state-of-the-art methods [8], [28], [33], [41], [43], [45], [46], [47], [48] for image deraining in terms of two evaluation metrics on five benchmark datasets in Table 1. Our DAT model achieves the best performance on all datasets, which indicates the robustness and effectiveness of our proposed model across different datasets. It is reasonable that our DAT significantly outperforms sequential modeling-based methods, RESCAN [28] and PReNet [48], since the self-attention mechanism is able to more sufficiently capture contextual information while learning degradation patterns. Compared to the convolution-based frameworks such as MPRNet [43], our DAT enjoys superior degradation modeling capability, benefiting from the proposed degradation-aware Transformer block which introducing distinguishable representations of rainy degradation patterns thorough contrastive learning.

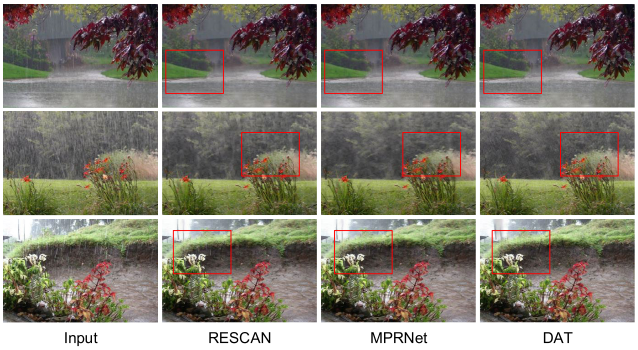
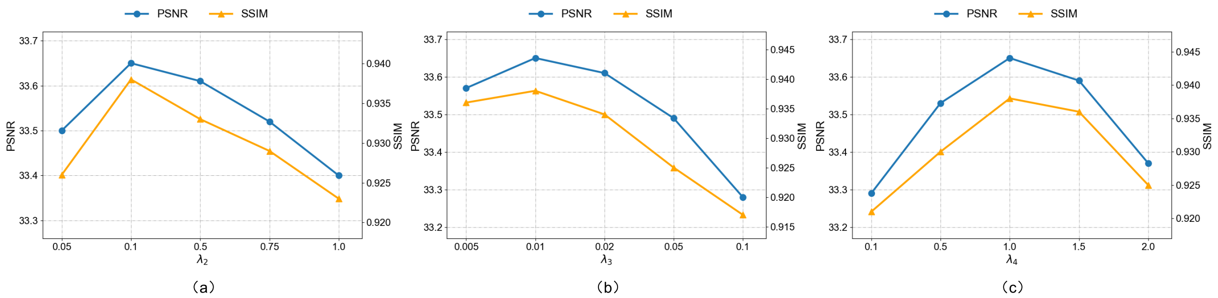


FIGURE 6. **Visual comparison for generalization performance on real-world rainy images by different methods.**

2) QUALITATIVE COMPARISON

We also perform a visual comparison with other state-of-the-art methods for single image deraining in Figure 2. The visual comparison demonstrates that our *DAT* restores the cleanest background image from the rainy image. By contrast, other competing methods for single image deraining still leave visible rain constituents or artifacts in the restored results, which reveals the superiority of rain pattern modeling by our proposed multi-level contrastive learning. For instance, in the fourth row in Figure 2, our *DAT* is able to thoroughly remove rain constituents, and reconstruct higher-quality background



(a) (b) (c)

FIGURE 7. Investigation on different weights of loss functions.

image with almost no obvious rain residues. It shows the advantageous performance of both rain separation and background reconstruction by our *DAT*.

3) GENERALIZATION PERFORMANCE

Besides performance evaluation on benchmark datasets for single image deraining, we also evaluate the generalization performance of different methods on real-world rainy images. As presented in Figure 6, our *DAT* consistently restore the highest quality results. This demonstrates again the effectiveness of our proposed degradation-aware Transformer model for single image deraining, especially the developed self-attention mechanism which leverages the discriminative degradation representations learned from the latent feature space.

TABLE 2. User study on 100 deraining results of Rain800 and Rain1400. 50 human subjects are performed for comparison between our *DAT* and three state-of-the-art methods for single image deraining.

Method	Share of the vote	
	Rain800	Rain1400
RESCAN [28]	0.6%	0.32%
MSPFN [41]	2.68%	3.96%
MPRNet [43]	15.16%	11.32%
DAT	81.56%	84.4%

4) USER STUDY

The quantitative comparison has their bias for the accurate evaluation of restoration performance. Thus, we perform user study on three competing state-of-the-art methods for single image deraining to avoid this bias, including RESCAN [28], MSPFN [41], and MPRNet [43]. There are 50 images that are randomly selected from the test datasets, and their results are sent to 50 human judges for ranking the restoration effectiveness of different methods. The results of user study on Rain800 and Rain1400 datasets are listed in Table 2, and our *DAT* achieves the highest votes than competing methods for single image deraining. This again reveals the superior advantages of our proposed degradation-aware deraining model.

C. ABLATION STUDY

To investigate the effectiveness of each proposed component in our *DAT*, we conduct experiments for ablation study. Specifically, we consider four variants of *DAT*: 1) **ConvNet**, which is a basic encoder-decoder based convolutional network; 2) **Base-T**, which leverages the transformer block of Swin-Transformer [49] to capture contextual information while modeling rain patterns and follows a U-shaped encoder-decoder framework; 3) **Conv-T**, which further introduces depth-wise and point-wise convolution in Base-T to preserve spatial structures and improve the quality of restored images; 4) **DAT**, which is the complete model of our method, which further leverages the proposed multi-level contrastive learning to perform degradation-aware self-attention in the proposed deraining Transformer.

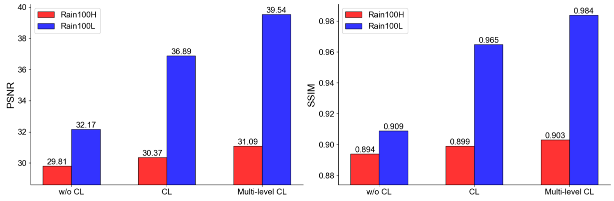


FIGURE 8. Ablation study on different way of using contrastive learning.

TABLE 3. Ablation study on our *DAT* in terms of PSNR and SSIM to investigate the effectiveness of each proposed technique in our model.

Method	Real100L		Rain100H	
	PSNR	SSIM	PSNR	SSIM
ConvNet	26.79	0.812	24.01	0.795
Base-T	29.95	0.886	28.15	0.860
Conv-T	32.17	0.909	29.81	0.894
DAT	**39.54**	**0.984**	**31.09**	**0.903**

1) EFFECT OF MULTI-LEVEL CONTRASTIVE LEARNING

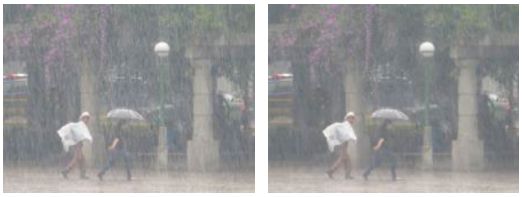
To investigate the effect of proposed multi-level contrastive learning, we conduct experiments on two benchmark datasets by different way of using contrastive learning. As presented in Figure 8, the results prove that the superiority of our multi-level contrastive learning, leading to a higher performance by more discriminative latent representations from the degradation-aware model.

TABLE 4. Model complexity of our *DAT* and two state-of-the-art methods for image reflection removal in terms of GFLOPs and Inference time.

Model	Params	GFLOPs	Inference time
MSPFN [41]	21.81M	186.3	0.120s
DMGN [47]	38.27M	232.5	0.066s
MPRNet [43]	20.14M	565.8	0.098s
DAT	18.96M	177.2	0.061s

2) EFFECT OF DEGRADATION-AWARE MULTI-HEAD SELF-ATTENTION MECHANISM

In Table 3, the significant performance gain from Conv-T to DAT demonstrates the effectiveness of proposed degradation-aware multi-head self-attention mechanism. This is reasonable since discriminative degradation representations are able to facilitating sufficiently modeling rain patterns by the self-attention mechanism of deraining Transformer.



(a) Input (b) DAT (ours)

FIGURE 9. Failure case.

3) INVESTIGATION ON THE LEVEL NUMBER OF MULTI-LEVEL CONTRASTIVE LEARNING

To explore how many levels of multi-level contrastive learning are optimal to restore rainy images, we train our DAT model with different levels of multi-level contrastive learning on Rain100H dataset. The experimental results are illustrated in Table 5, which shows deploying 3 levels achieves the best performance.

TABLE 5. Ablation study of level number in multi-level contrastive learning.

L	0	1	2	3	4
PSNR	29.81	30.37	30.65	31.09	30.83
SSIM	0.894	0.899	0.902	0.903	0.901

4) EFFICIENCY ANALYSIS

Besides performance analysis, we also compare the computational cost of *DAT* with other competing state-of-the-art methods for single image deraining in Table 4. The results show that the performance gain by our *DAT* results from effective model design rather than increased parameters.

5) HYPER-PARAMETER SELECTION

To determine the weight of different loss functions, we conduct experiments on Rain100L dataset with different weights. The results in Figure 7 show that our *DAT* achieves the best performance while setting λ_1=1, λ_2=0.1, λ_3=0.01, and λ_4=1.

6) FAILURE CASES

Even if our *DAT* has shown significant performance while restoring rainy images, negative results are observed for a few challenging cases as well. As shown in Figure 9, our *DAT* fails to effectively restore high-quality background image when degrdations are mixed, such as haze and rain streaks. In our future work, we plan to solve this problem.

V. CONCLUSION

In this work, we propose a Degradation-Aware Transformer (DAT) for single image deraining, which employs multi-level contrastive learning to learn a discriminative latent feature space, and thus obtain degradation-aware representations. To this end, we design a Degradation-Aware Transformer Block (DATB) that performs self-attention mechanism under the guidance of discriminative degradation representation while capturing long-term contextual relations between pixels. Therefore, DAT is able to more completely model diverse rain patterns and restore higher-quality background image. We have conducted extensive experiments on various benchmark datasets for image rain streak removal, and the results show that our DAT is able to consistently outperform other state-of-the-art methods both qualitatively and quantitatively.

REFERENCES

[1] H. Zhang and V. M. Patel, "Convolutional sparse and low-rank coding-based rain streak removal," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 1259–1267.

[2] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3397–3405.

[3] Y.-L. Chen and C.-T. Hsu, "A generalized low-rank appearance model for spatio-temporally correlated rain streaks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1968–1975.

[4] Y. Luo and J. Ling, "Single-image de-raining using low-rank matrix approximation," *Neural Comput. Appl.*, vol. 32, no. 11, pp. 7503–7514, Jun. 2020.

[5] T. Liu, H. Tang, D. Zhang, S. Zeng, B. Luo, and Z. Ai, "Feature-guided dictionary learning for patch-and-group sparse representations in single image deraining," *Appl. Soft Comput.*, vol. 113, Dec. 2021, Art. no. 107958.

[6] H. Wang, Q. Xie, Q. Zhao, Y. Li, Y. Liang, Y. Zheng, and D. Meng, "RCDNet: An interpretable rain convolutional dictionary network for single image deraining," 2021, *arXiv:2107.06808*.

[7] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2516–2525.

[8] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1715–1723.

[9] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8014–8023.

[10] C. Wang, Y. Wu, Z. Su, and J. Chen, "Joint self-attention and scale-aggregation for self-calibrated deraining network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2517–2525.

[11] H.-H. Yang, C. H. Yang, and Y. F. Wang, "Wavelet channel attention module with a fusion network for single image deraining," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 883–887.

[12] K. Jiang, Z. Wang, P. Yi, C. Chen, Z. Han, T. Lu, B. Huang, and J. Jiang, "Decomposition makes better rain removal: An improved attention-guided deraining network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3981–3995, Oct. 2021.

[13] Z. Guo, M. Hou, M. Sima, and Z. Feng, ''DerainAttentionGAN: Unsupervised single-image deraining using attention-guided generative adversarial networks,'' *Signal, Image Video Process.*, vol. 16, no. 1, pp. 185–192, Feb. 2022.

[14] A. Vaswani, N. Shazeer, and N. Parmar, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[15] G. Wang, C. Sun, and A. Sowmya, ''Context-enhanced representation learning for single image deraining,'' *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1650–1674, May 2021.

[16] Q. Yi, J. Li, Q. Dai, F. Fang, G. Zhang, and T. Zeng, ''Structure-preserving deraining with residue channel prior guidance,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4218–4227.

[17] W.-Y. Hsu and W.-C. Chang, ''Recurrent wavelet structure-preserving residual network for single image deraining,'' *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109294.

[18] T. Chen, S. Kornblith, and M. Norouzi, ''A simple framework for contrastive learning of visual representations,'' in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, ''Momentum contrast for unsupervised visual representation learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[20] O. Henaff, ''Data-efficient image recognition with contrastive predictive coding,'' in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.

[21] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma, ''Contrastive learning for compact single image dehazing,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10546–10555.

[22] H. Ji, X. Feng, W. Pei, J. Li, and G. Lu, ''U2-former: A nested U-shaped transformer for image restoration,'' 2021, *arXiv:2112.02279*.

[23] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, ''Automatic single-image-based rain streaks removal via image decomposition,'' *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1742–1755, Apr. 2012.

[24] L.-J. Deng, T.-Z. Huang, X.-L. Zhao, and T.-X. Jiang, ''A directional global sparse model for single image rain removal,'' *Appl. Math. Model.*, vol. 59, pp. 662–679, Jul. 2018.

[25] J.-H. Kim, C. Lee, J.-Y. Sim, and C.-S. Kim, ''Single-image deraining using an adaptive nonlocal means filter,'' in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 914–917.

[26] Y. Wang, S. Liu, C. Chen, and B. Zeng, ''A hierarchical approach for rain or snow removing in a single color image,'' *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3936–3950, Aug. 2017.

[27] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, ''Rain streak removal using layer priors,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2736–2744.

[28] X. Li, J. Wu, and Z. Lin, ''Recurrent squeeze-and-excitation context aggregation net for single image deraining,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 254–269.

[29] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, ''Generative adversarial nets,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[30] D. P. Kingma and M. Welling, ''Auto-encoding variational Bayes,'' 2013, *arXiv:1312.6114*.

[31] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, ''Deep joint rain detection and removal from a single image,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1685–1694.

[32] G. Li, X. He, W. Zhang, H. Chang, L. Dong, and L. Lin, ''Non-locally enhanced encoder–decoder network for single image de-raining,'' in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1056–1064.

[33] H. Zhang and V. M. Patel, ''Density-aware single image de-raining using a multi-stream dense network,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 695–704.

[34] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Jr., J. Zhang, X. Guo, and X. Cao, ''Single image deraining: A comprehensive benchmark analysis,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3833–3842.

[35] T. Park, A. A. Efros, and R. Zhang, ''Contrastive learning for unpaired image-to-image translation,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 319–345.

[36] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, ''Unsupervised degradation representation learning for blind super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10576–10585.

[37] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, ''Image-to-image translation with conditional adversarial networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[39] J. Canny, ''A computational approach to edge detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, ''Perceptual losses for real-time style transfer and super-resolution,'' in *Proc. Eur. Conf. Comput. Vis.*, Springer, Cham 2016, pp. 694–711.

[41] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, and J. Jiang, ''Multi-scale progressive fusion network for single image deraining,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8343–8352.

[42] H. Zhang, V. Sindagi, and V. M. Patel, ''Image de-raining using a conditional generative adversarial network,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3943–3956, Nov. 2020.

[43] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, ''Multi-stage progressive image restoration,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14816–14826.

[44] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[45] H. Wang, Q. Xie, Q. Zhao, and D. Meng, ''A model-driven deep neural network for single image rain removal,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3100–3109.

[46] X. Feng, H. Ji, B. Jiang, W. Pei, F. Chen, and G. Lu, ''Contrastive feature decomposition for image reflection removal,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[47] X. Feng, W. Pei, Z. Jia, F. Chen, D. Zhang, and G. Lu, ''Deep-masking generative network: A unified framework for background restoration from superimposed images,'' *IEEE Trans. Image Process.*, vol. 30, pp. 4867–4882, 2021.

[48] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, ''Progressive image deraining networks: A better and simpler baseline,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3932–3941.

[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[50] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, *arXiv:1704.04861*.

[51] J. L. Ba, J. R. Kiros, and G. E. Hinton, ''Layer normalization,'' 2016, *arXiv:1607.06450*.

PEIJUN ZHAO received the bachelor's degree in electronic information engineering from Xinyang Normal University, in 2006, and the master's degree in pattern recognition and intelligent system from Zhengzhou University, Zhengzhou, China, in 2010. Since 2019, she has been a Lecturer with Xinyang Agriculture and Forestry University. Her research interests include fault diagnosis and fault tolerant control, pattern recognition, and intelligent systems.

TONGJUN WANG received the bachelor's degree in electronic information engineering from Xinyang Normal University, in 2006, and the master's degree in control theory and control engineering from Zhengzhou University, Zhengzhou, China, in 2009. Since 2021, he has been an Associate Professor with Xinyang Agriculture and Forestry University. His research interests include environment intelligent perception and high-speed information processing.