Hybrid CNN-Transformer Feature Fusion for Single Image Deraining

Xiang Chen[1], Jinshan Pan[1]*, Jiyang Lu[2], Zhentao Fan[2], Hao Li[1]

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

Abstract

Since rain streaks exhibit diverse geometric appearances and irregular overlapped phenomena, these complex characteristics challenge the design of an effective single image deraining model. To this end, rich local-global information representations are increasingly indispensable for better satisfying rain removal. In this paper, we propose a lightweight Hybrid CNN-Transformer Feature Fusion Network (dubbed as HCT-FFN) in a stage-by-stage progressive manner, which can harmonize these two architectures to help image restoration by leveraging their individual learning strengths. Specifically, we stack a sequence of the degradation-aware mixture of experts (DaMoE) modules in the CNN-based stage, where appropriate local experts adaptively enable the model to emphasize spatially-varying rain distribution features. As for the Transformer-based stage, a background-aware vision Transformer (BaViT) module is employed to complement spatially-long feature dependencies of images, so as to achieve global texture recovery while preserving the required structure. Considering the indeterminate knowledge discrepancy among CNN features and Transformer features, we introduce an interactive fusion branch at adjacent stages to further facilitate the reconstruction of high-quality deraining results. Extensive evaluations show the effectiveness and extensibility of our developed HCT-FFN. The source code is available at https://github.com/cschenxiang/HCT-FFN.

Introduction

Single image deraining (SID) is the task of recovering clear and rain-free background from the given rainy images, since the images captured under rainy conditions significantly degrade the performance of downstream computer vision systems (including autonomous driving and video surveillance, etc.), which has drawn widespread attention in recent years.

Early prior-based methods (Kang, Lin, and Fu 2011; Luo, Xu, and Ji 2015; Li et al. 2016; Zhang and Patel 2017) attempt to remove the rain by relying on statistical properties of rain components and clear backgrounds. However, these hand-crated priors from human observation may not always hold in case of the complex and varying rainy scenarios.

To circumvent hypothetical priors dependency, numerous CNN-based networks (Yang et al. 2020; Yu et al. 2022) have been proposed for SID, which achieves remarkable progress thanks to the rapid growing complicated architectures (Jiang et al. 2020; Zamir et al. 2021; Mou, Wang, and Zhang 2022) and learning strategies (Zhou et al. 2021; Xiao et al. 2021; Chen et al. 2022). However, these approaches still encounter performance bottlenecks due to the local receptive fields of the CNN-based operations, which limits the ability to cope with long-range dependency information. To this end, recent Transformers have emerged in computer vision field (Dosovitskiy et al. 2020; Chen et al. 2021), which is attributed to the unique advantage of self-attention with global feature interaction. Since then, several modified Transformer-based architectures (Xiao et al. 2022; Wang et al. 2022; Zamir et al. 2022) have also been developed for SID task achieving superior performance over previous CNN-based models.

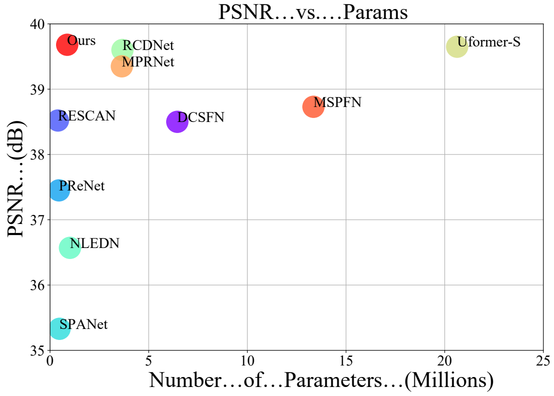


Figure 1: Comparison results on the Rain100L dataset. Our method not only reconstructs a high-quality output but also achieves the best performance-parameter trade-off.

Since the rain streak layer and rain-free background layer are highly interlaced, global and local representation learning are equally important for the challenging SID task, while the self-attention in Transformer does not manipulate the local invariance that CNNs do well. Afterwards, some studies (Yuan et al. 2021; Wu et al. 2021) attempt to introduce convolutional operations into vision Transformers, but they do not play a full role for low-level image restoration. To mitigate this problem, a few recent works (Guo et al. 2022; Jiang

*Corresponding author

et al. 2022) try to combine these two structures to construct a hybrid model aiming to inherit advantages of CNN and Transformer. This naturally raises a crucial question: how to effectively integrate both CNN features and Transformer features? In fact, an intuitive observation is that there are indeterminate knowledge discrepancies among convolution-based CNN features and self-attention-based Transformer features (Park and Kim 2022), thus simply concatenating or adding these features is inefficient for significant performance gain. Therefore, it is of great interest to tailor design fusion models so that they can better facilitate rain removal.

In this work, we present a new hybrid network that combines the features by CNN and Transformer for comprehensive rain distribution prediction, which is expected to produce better deraining results than any individual model. Following the past successful inspired designs (Ren et al. 2019; Jiang et al. 2020), we specifically formulate our fusion framework in a stage-by-stage progressive fashion due to the complexity of SID. To alleviate the learning difficulty, we propose to separately extract the intra-stage hierarchical feature via a backbone branch and adaptively aggregate the inter-stage complementary feature via an auxiliary branch. As such, all the stage representations are richer.

Specifically, the body of Hybrid CNN-Transformer Feature Fusion Network (HCT-FFN) consists of three sequential stages, which can excavate the useful information from previous stage to guide the later stage. In the backbone branch, CNN-based backbone is applied in the first and last stages, while Transformer-based backbone is used in the intermediate stage. In terms of the CNN-based stage, we stack a series of degradation-aware mixture of experts (DaMoE) modules that adaptively restore an image degraded by the spatially-varying rain distribution. By doing so, experts (local CNN operations in parallel) are able to focus on assigning corresponding intensity weights for different degradation factors depending on the inputs, so that we can facilitate the model to adaptively remove rainy effects of different appearances. As for the Transformer-based stage, a background-aware vision Transformer (BaViT) module is employed to eliminate the spatially-long rain degradation by modeling long-range dependencies, since the multi-head self-attention facilitates global texture and structure recovery. Instead of simply concatenating the features of two adjacent stages, we also introduce an interactive fusion branch (IFB) to encode the inter-stage correlation among backbone features and reconstruction features. In this way, IFB can allow to explore complementary components of hybrid features by CNN and Transformer from each other through stage-wise reconstruction for further refinement. Finally, comprehensive experiments show that our hybrid fusion model achieves the best performance-parameter trade-off, as shown in Fig. 1.

Our main contributions are summarized as follows:

- We propose an end-to-end hybrid model for SID, HCT-FFN, integrating the intra-stage advantages of CNN and Transformer paradigms to achieve a strong deraining baseline in a stage-by-stage unified architecture.
- We show that the inter-stage interactive fusion can alleviate the knowledge discrepancy among the features by CNN and Transformer, in order to better facilitate rain removal.
- We perform extensive experiments to demonstrate the effectiveness and extensibility of the proposed HCT-FFN.

Related Work

Single Image Deraining

Early deep CNN-based networks (Yang et al. 2017; Fu et al. 2017a; Zhang and Patel 2018; Li et al. 2018) have emerged for SID as a better option compared to hand-crafted priors. By further optimizing the network structure, researchers employ the recursive computation (Ren et al. 2019; Zamir et al. 2021) or the multi-scale representation (Yasarla and Patel 2019; Jiang et al. 2020) to effectively produce rain-free results. Instead of prevailing CNN-based pipeline, Transformer is recently introduced as a new network backbone to account for performance gain. For low-level image restoration, typical architectures include IPT (Chen et al. 2021), Restormer (Zamir et al. 2022), and Uformer (Wang et al. 2022). However, most of these methods blindly stack pure Transformer-based components to replace original CNNs, which inevitably generates high computational cost leading to a bloated model with an excessive amount of parameters. For instance, concurrent Restormer (Zamir et al. 2022) requires 26.10 Million parameters to obtain competitive results. Few attempts have been made to fully consider complementary merits between CNN and Transformer, thus difficult to enable the model to offer the optimal balance between size and performance. More recently, ELF (Jiang et al. 2022) is first presented to unify these two architectures into an association learning-based lightweight hybrid deraining model. Different from it, inspired by the progressive learning-based formulation, we are committed to design a new hybrid deraining network by gradually removing rain streaks in a stage-by-stage manner.

Vision Transformer

Transformer-based models (Vaswani et al. 2017) originally bring significant breakthroughs to the natural language processing (NLP) field. Benefiting from the powerful capability in modeling long-range information with the help of the self-attention mechanism, the birth of Vision Transformer (ViT) (Dosovitskiy et al. 2020) makes computer vision community shine again, which has witnessed prominent improvements among high-level vision tasks (Carion et al. 2020; Liu et al. 2021; Zheng et al. 2021). Likewise, recent studies have applied variants of ViT in a host of low-level vision problems and opened up a new perspective, such as image dehazing (Guo et al. 2022), and image super-resolution (Gao et al. 2022). Furthermore, in terms of recent hybrid models, information fusion between Transformer features and CNN features has become a key step. In (Guo et al. 2022), these features are aggregated by learning the modulation matrices to solve the feature inconsistency issue. However, these fusions lack interactivity because only limited intra-stage connections are considered. In this paper, we propose to explore the inter-stage interactive fusion to guide image reconstruction by encoding the correlation among these two joint features.

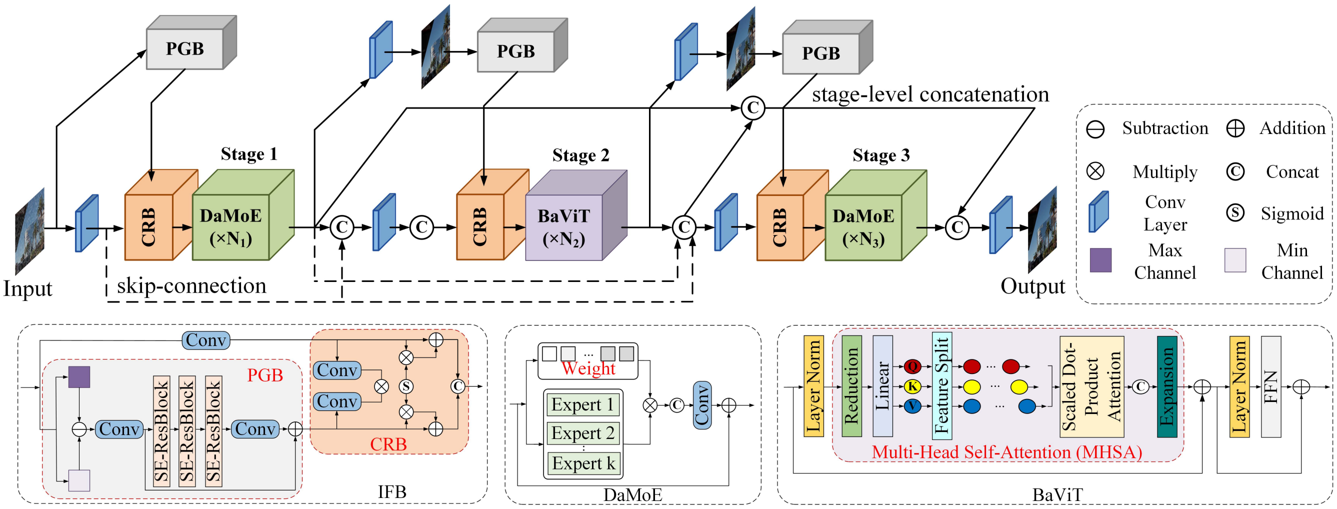


Figure 2: The overall framework of the proposed Hybrid CNN-Transformer Feature Fusion Network (HCT-FFN), which mainly contains (1) degradation-aware mixture of experts (DaMoE) module, (2) background-aware vision Transformer (BaViT) module, and (3) interactive fusion branch (IFB) with prior guidance block (PGB) and coupled representation block (CRB).

Proposed Method

This section mainly introduces the proposed end-to-end Hybrid CNN-Transformer Feature Fusion Network (HCT-FFN) to remove undesirable rain streaks in a stage-by-stage manner. The whole framework is illustrated in Fig. 2, which contains three recursive stages. In the first and last CNN-based stages, we stack the degradation-aware mixture of experts (DaMoE) modules as the network backbone to extract local features for spatially-varying rain degradation. Meanwhile, a background-aware vision Transformer (BaViT) module is employed as the backbone of the intermediate Transformer-based stage to capture global dependencies for spatially-long rain appearance. Furthermore, prior guidance block (PGB) and coupled representation block (CRB) are incorporated in interactive fusion branch (IFB) to further provide complementary information for the model, so that high-quality clear outputs can be gradually reconstructed. To enable the model to learn richer features during image restoration process, we fuse the output features of the previous stage with the output features of the current stage using the skip-connection and stage-level concatenation. With this design, useful information from the previous stage can be fully excavated to guide the later stage, allowing the redundant feature to deep layers without too much processing, thus selectively focusing on more important information. In what follows, we will describe the details about the above-mentioned components.

Degradation-aware Mixture of Experts

In the CNN-based stages, DaMoE module is the key part to successfully restore complicated rain distribution. Considering the design of recent effective CNN models (Suganuma, Liu, and Okatani 2019), we elaborately select multiple local CNN operations to form parallel layers, dubbed as *experts*, which involve a average pooling with receptive field of 3×3, separable convolution layers with kernel sizes of 1×1, 3×3, 5×5, 7×7, and dilated convolution layers with kernel sizes of 3×3, 5×5, 7×7. Different from the conventional mixture of experts (Jacobs et al. 1991; Ren et al. 2018), our DaMoE module does not attach an external gating network. Instead, we make the self-attention scheme (Hu, Shen, and Sun 2018; Kim, Ahn, and Sohn 2020) become a switcher of different experts to adaptively select the importance of diverse representations depending on the inputs, which will collaboratively help context aggregation. Given an input feature map $\mathbf{x}_c \in R^{C\times H\times W}$, we first apply the channel-wise average to generate C-dimensional channel descriptor $\mathbf{z}_c \in R^C$:

$$\mathbf{z}_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{x}_c(i,j), \qquad (1)$$

where $\mathbf{x}_c(i,j)$ is the (y,x) position of the feature $\mathbf{x}_c$. Then, the attention weight vector of each expert is allocated corresponding to the learnable weight matrices $W_1 \in R^{T\times C}$ and $W_2 \in R^{O\times T}$. Here, T is the dimension of the weight matrices. To avoid altering the sizes of its inputs and outputs, we zero pad the input feature maps computed by each expert. With this formulation, the extracted degradation features by feeding a DaMoE module denotes F_{deg}, then we have

$$F_{deg} = [f_{exp}^{O}(W_2\sigma(W_1\mathbf{z}))],\ \text{for } O=1,2,\ldots,k \qquad (2)$$

where f_{exp} and O represent the expert layer and the number of experts respectively. $\sigma(\cdot)$ is a ReLU function, and $[\cdot]$ denotes the channel-wise concatenation. Note that we employ a skip connection between each DaMoE module to bridge across continuous intermediate features for stable training.

Finally, the output of the N-th DaMoE module is calculated by

$$F_{DaMoE}^{N} = f_{1\times1}(F_{deg}) + F_{DaMoE}^{N-1}, \qquad (3)$$

where $f_{1\times1}(\cdot)$ denotes a convolutional layer with C filters.

Background-aware Vision Transformer

In the Transformer-based stage, we introduce a BaViT module to help accurate background recovery, thanks to the advantage of Transformer in learning long-range dependencies within the global information. Unlike ViT (Dosovitskiy et al. 2020), we first employ the *unfold* operation to split the input feature maps F_{in} into $H \times W$ patches $F_{in}^{*} \in R^{k \times k \times C}$ by a $k \times k$ kernel. Intuitively, this pre-processing step naturally reflects the position information of each patch. Following that, these patches are sent directly to the body of BaViT module. Mathematically, the encoding procedures are expressed as

$$F_{mid} = F_{in}^{*} + f_{MHSA}\left(f_{Norm}\left(F_{in}^{*}\right)\right), \quad (4)$$

$$F_{BaViT} = F_{mid} + f_{FFN}\left(f_{Norm}\left(F_{mid}\right)\right), \quad (5)$$

where $f_{MHSA}(\cdot)$ and $f_{FFN}(\cdot)$ denote the multi-head self-attention (MHSA) and feed-forward network (FFN), respectively. $f_{Norm}(\cdot)$ refers to the layer normalization operation. Finally, we use the *fold* operation to reconstruct feature maps F_{BaViT} of the BaViT module.

As shown in Fig. 2, MHSA and FFN are key ingredients for Transformer, which aim to perform interaction and transformation between tokens. Specifically, in the MHSA part, we first halve the number of channels using a reduction layer and then project the input embedding to the Q (query), K (keys), and V (values) elements through a linear layer. To enrich the background representation, multi-head attention (Vaswani et al. 2017) is performed on Q, K and V. Inspired by (Lu et al. 2022), we adopt *feature split* operation to divide Q, K, V into s equal segments along the channel dimension to obtain $\{Q_1, Q_2, \ldots, Q_s\}$, $\{K_1, K_2, \ldots, K_s\}$, and $\{V_1, V_2, \ldots, V_s\}$. For each segment, it has $C_k = \frac{C}{S}$ channels. Each triplet of these segments is usually calculated by scaled dot-product attention function:

$$f_{sdpa} = \text{softmax}\left(\frac{Q_i K_i^{\top}}{\sqrt{C_k}}\right) V_i, \text{ for } i = 1, 2, \ldots, s. \quad (6)$$

Lastly, we concatenate all the output of multi-head attention, and then utilize an extension layer to recover the number of channels. To keep the block simple, the FFN part is composed of two Multi-Layer Perceptions (MLPs) layers.

Interactive Fusion Branch

To refine the inter-stage representation among CNN features and Transformer features, we formulate an IFB to provide additional complementary information to the backbone network. Compared to direct concatenating the features of two adjacent stages, our IFB tends to be more flexible and effective. Specifically, we first make full use of image priors to embed into IFBs as feature guidances, thus facilitating the stage-wise reconstruction of high-quality results. Similar to (Li, Tan, and Cheong 2018; Yi et al. 2021), residual channel prior (RCP) is applied due its advantage in extracting clear object structure. It is regarded as the residual result of the maximum and minimum channel values of the rainy image, which is calculated without any additional parameters:

$$F_{rcp}(x) = \max_{c \in \{R,G,B\}} I^{c}(x) - \min_{d \in \{R,G,B\}} I^{d}(x). \quad (7)$$

Following it, we adopt SE-Resblocks (Hu, Shen, and Sun 2018) to further enhance channel-wise feature propagation. Formally, the final output F_{pgb} of PGB is defined by

$$F_{pgb} = f_{3\times3}\left(f_{SE}^{3}\left(f_{3\times3}\left(F_{rcp}\right)\right)\right) + f_{3\times3}\left(F_{rcp}\right), \quad (8)$$

where $f_{SE}^{3}(\cdot)$ is the cascade feature of three SE-Resblocks.

Then, PGB is fed into CRB to encode the mixture relations among backbone features and reconstruction features, which can learn redundant components adaptively from each other for further refinement. Here, two convolution layers with a kernel size of 3×3 are used to map the backbone feature F_{bac} from the output of previous stage (*i.e.*, F_{DaMoE}^{N} or F_{BaViT}) and reconstructed image feature F_{pgb} from PGB, respectively. Next, we use element multiplication to calculate the similarity map S between two branch features:

$$S = \text{sigmoid}\left(f_{3\times3}\left(F_{bac}\right) \otimes f_{3\times3}\left(F_{pgb}\right)\right), \quad (9)$$

where $\otimes$ represents pixel-wise product.

Lastly, the original features are further added to the activated features, and the summed features are concatenated to return a refined joint representation F_{crb} of CRB:

$$F_{crb} = \text{concat}\left(S \otimes F_{bac} + F_{bac}, S \otimes F_{pgb} + F_{pgb}\right). \quad (10)$$

With the help of this interactive fusion pattern architecture, we can not only fully utilize the dependencies of deep features across stages, but also boost the collaborative representation from CNN and Transformer to help image restoration.

Loss Function

To supervise the learning process of the network, we choose two proper loss functions as training objectives to drive the model optimization. Mean squared error (MSE) loss (Zhang and Patel 2018) is widely adopted to compute the pixel-level difference between the recovered image B_i and corresponding ground truth B, expressed as follows:

$$\mathcal{L}_{mse} = \frac{1}{HWC} \sum_{x=1}^{H} \sum_{y=1}^{W} \sum_{z=1}^{C} \left\| B_i - B \right\|^2, i = 1, 2, 3 \quad (11)$$

where i denotes different stages, and H, W and C are height, width and number of channels, respectively.

To further improve the deraining results with high fidelity, we consider the structural similarity (SSIM) to compare the structural differences, which is calculated as follows:

$$SSIM(B_i, B) = \frac{2\mu_{B_i}\mu_B + C_1}{\mu_{B_i}^2 + \mu_B^2 + C_1} \cdot \frac{2\sigma_{B_i B} + C_2}{\sigma_{B_i}^2 + \sigma_B^2 + C_2}, \quad (12)$$

where μ_{B_i} and μ_B are the average of B_i and B over pixels, σ_{B_i} and σ_B are the variances of B_i and B, $\sigma_{B_i B}$ is the covariance between B_i and B. C_1 and C_2 are two fixed constants. Then, the negative SSIM loss (Ren et al. 2019) for recovered image is given by:

$$\mathcal{L}_{ssim} = 1 - SSIM(B_i, B). \quad (13)$$

Finally, the overall loss $\mathcal{L}_{total}$ for training our network is the combination of the above two losses as:

$$\mathcal{L}_{total} = \mathcal{L}_{mse} + \lambda \mathcal{L}_{ssim}, \quad (14)$$

where the coefficient λ is empirically set to 0.2 for balancing each loss term.

Datasets			Synthetic						Real-world	
			Rain100L		Rain100H		Rain12		RainDS-RS100	
Methods		Param	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Rainy Input		-	26.90	0.838	13.56	0.370	30.14	0.855	23.58	0.651
Prior-based methods	GMM (CVPR'16)	-	29.05	0.871	15.23	0.449	32.14	0.914	23.73	0.559
	JCAS (CVPR'17)	-	28.54	0.852	14.62	0.451	33.10	0.930	24.04	0.556
CNN-based methods	DNN (CVPR'17)	0.06	32.38	0.925	22.85	0.725	34.04	0.933	24.61	0.681
	RESCAN (ECCV'18)	0.15	38.52	0.981	29.62	0.872	36.43	0.951	25.84	0.686
	PReNet (CVPR'19)	0.17	37.45	0.979	30.11	0.905	36.66	0.961	26.29	0.718
	JORDER_E (TPAMI'19)	-	38.59	0.983	30.50	0.896	36.69	0.962	26.48	0.715
	MSPFN (CVPR'20)	13.35	38.73	0.978	30.63	0.898	36.85	0.957	26.55	0.721
	MPRNet (CVPR'21)	3.63	39.45	0.982	30.92	0.904	37.26	0.960	26.86	0.725
Transformer-based methods	IPT (CVPR'21)	115.5	**41.62**	**0.988**	-	-	-	-	-	-
	Uformer-B (CVPR'22)	50.8	39.76	0.983	31.06	0.908	37.10	0.958	26.83	0.728
Hybrid-based methods	ELF (MM'22)	1.53	36.67	0.968	30.48	0.896	-	-	-	-
	Ours	0.87	39.70	0.985	**31.51**	**0.910**	**37.54**	**0.963**	**27.02**	**0.734**

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

Experiments

Experimental Settings

Datasets. We conduct deraining experiments on four public rain streak datasets, including Rain100L (Yang et al. 2017), Rain100H (Yang et al. 2017), Rain12 (Li et al. 2016), and RainDS-Real (Quan et al. 2021). With light and heavy types of synthetic rain streaks, Rain100L and Rain100H contain 1,800 image pairs for training and 100 image pairs for testing. Rain12 contains 12 light rainy images. Based on autonomous driving scenario, (Quan et al. 2021) release various image pairs corrupted by raindrops and rain streaks, which consists of two subsets, RainDS-Syn and RainDS-Real. As we mainly focus on removing rain streaks, we only adopt a subset of RainDS-Real, named RainDS-RS100, where 150 real-world rainy images are chosen as training data and the other 100 pairs are selected for testing. In addition, we also randomly choose 20 real rainy images without ground truths from Internet-Data (Wang et al. 2019; Yang et al. 2020) as the evaluation of generalization performance.

Comparison methods. We compare our method with two prior-based algorithms (i.e., GMM (Li et al. 2016) and JCAS (Gu et al. 2017)), six CNN-based appraohces (i.e., DDN (Fu et al. 2017b), RESCAN (Li et al. 2018), PReNet (Ren et al. 2019), JORDER_E (Yang et al. 2019), MSPFN (Jiang et al. 2020), and MPRNet (Zamir et al. 2021)), two Transformer-based networks (i.e., IPT (Chen et al. 2021) and Uformer-B (Wang et al. 2022)), and one hybrid-based model (i.e., ELF (Jiang et al. 2022)). Due to hard-ware constraints, IPT is only evaluated on Rain100L, and the corresponding results refer to their original paper. As the code of ELF is not available, we refer to some results presented in their paper. For other approaches, we retrain the models using the default settings provided by the authors if there are no pretrained models, otherwise we evaluate them with their online codes.

Evaluation metrics. Since the ground truths available, we adopt two commonly-used metrics for quantitative comparison, and they are Peak Signal to Noise Ratio (PSNR) (Huynh-Thu and Ghanbari 2008) and Structural Similarity (SSIM) (Wang et al. 2004). Following (Wang et al. 2020; Xiao et al. 2022), we calculate PSNR/SSIM metrics in Y channel of YCbCr space. For the rainy images without their clean labels, two popular non-reference indicators, Naturalness Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2012) and Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) (Mittal, Moorthy, and Bovik 2012), are employed for evaluating deraining performance.

Implementation details. The proposed network is implemented in PyTorch framework using Adam optimizer with a learning rate of 0.0001 to minimize $\mathcal{L}_{total}$ by 400 epochs. In our model, $\{N_1, N_2, N_3\}$ are set to $\{4, 3, 4\}$. During training, we run all of our experiments with batch size of 4 and patch size of 128 on one NVIDIA Tesla V100 GPU (32G). In the DaMoE module, we set $k = 8$ for the number of experts and $T = 32$ for the weight matrixs. Each convolutional layer has a $C = 16$ filter with stride of 1. In the BaViT module, we set $k = 3$ for the kernel size and $s = 4$ for splitting segment. The number of heads in MHSA is set to 8. For data augmentation, vertical and horizontal flips are randomly applied. The loss trade-off parameter is defined via cross validation using the validation set, and the whole pipeline is performed in an end-to-end fashion without costly large-scale pretraining (Chen et al. 2021).

Experimental Results

Synthetic datasets. Tab. 1 presents the quantitative results on different deraining benchmarks. Apparently, the proposed method significantly competes previous popular derainers on the Rain100H and Rain12 datasets, which reveals that our method can properly handle diverse types of spatially-varying rain streaks. And most remarkably, our designed HCT-FFN achieves prominent improvement in term of PSNR on the Rain100L and Rain100H benchmarks. Fig. 3 further shows visual comparison between samples generated by different baselines. It can be seen that Uformer-B is sensitive to local slender rain streaks. Besides, the results of pure CNN-based models are flawed in terms of global texture recovery. By contrast, our results are more consistent with that of the ground truths.

Real-world datasets. In order to further practical evalution

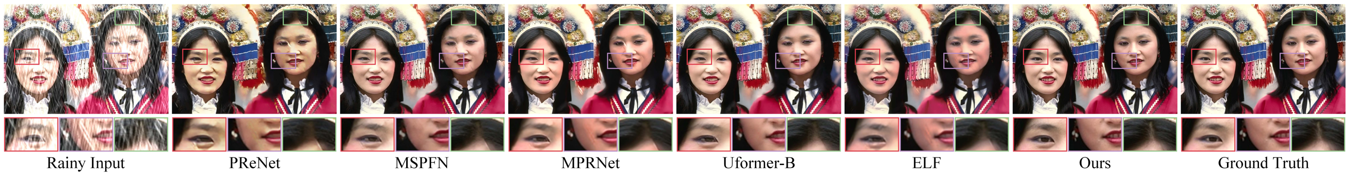


Figure 3: Visual comparison on the Rain100H dataset. Best viewed by zooming in the figures on high-resolution displays.

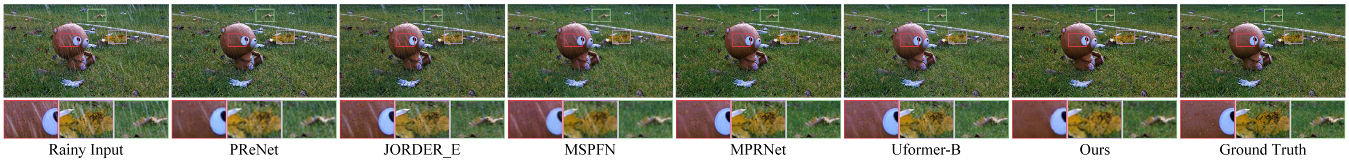


Figure 4: Visual comparison on the RainDS-RS100 dataset. Best viewed by zooming in the figures on high-resolution displays.

Method	PReNet	MSPFN	MPRNet	Uformer	Ours
NIQE	5.489	5.626	5.658	5.034	**4.743**
BRISQUE	33.576	42.159	37.195	33.293	**28.709**

Table 2: Comparison of quantitative results on Internet-Data dataset. Note that lower scores indicate better image quality.

Stage Number	Stage = 1	Stage = 2	Stage = 3
PSNR / SSIM	36.31 / 0.952	38.87 / 0.974	**39.70 / 0.985**

Table 3: Ablation analysis for different number of recursive stages. Stage = 1 also indicates that BaViT module is none.

in real-world rainy scenes, Tab. 2 and the last column of Tab. 1 compare the deraining results on the RainDS-RS100 dataset quantitatively. As expected, our developed method continues to achieve the highest PSNR/SSIM values and the lowest NIQE/BRISQUE scores, demonstrating the effectiveness and superiority of HCT-FFN, especially in the real rain with complicated rainy conditions. The reason behind is that our model enjoys the powerful abilities from the hybrid feature fusion of CNN and Transformer. Through the comparison in Fig. 4 and Fig. 6, our method successfully removes most rain perturbation and owns visual pleasant recovery results on several challenging exemples, which implies that it can well generalize to unseen real-world data types.

Ablation Studies

We study the individual components and parameter choices on the final deraining performance. Here, all ablation studies are conducted within the same training settings and environment using Rain100L dataset to ensure a fair comparison.

The Number of experts. To analyze the impact of different number of experts in each DaMoE module, we perform an experiment based on the parallel layer configuration in Fig. 5. When using single expert models, performance is dramatically degraded compared with multi-expert models. Unlike setting all experts to the same structure (Kim, Ahn, and Sohn 2020), our multi-expert structure is more diverse, which brings their own gains to the performance due to different receptive fields and disparate local CNN operations.

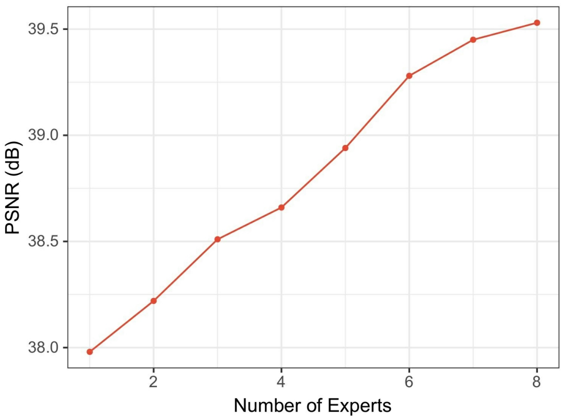
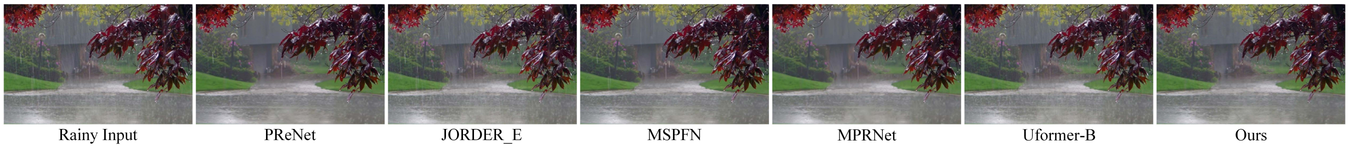


Figure 5: Ablation analysis for different number of experts in each DaMoE module.

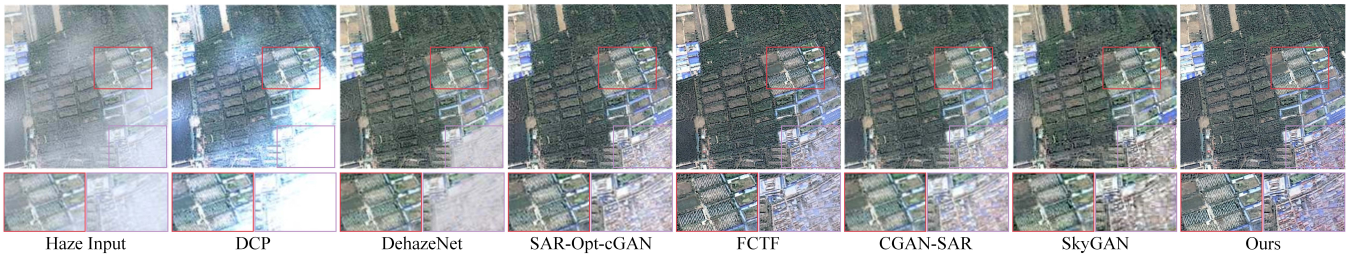
The Number of recursive stages. To analyze the effect about different number of recursive stage, Tab. 3 records the PSNR/SSIM of corresponding models. We can observe that: (1) BaViT module brings a great contribution to the baseline model, thanks to its advantage of modeling global interactions from non-local regions. (2) Stage-by-stage progressive learning can gradually eliminate the remaining rain streaks, thus achieving excellent deraining quality in the final stage.

Sequence of different stages. To analyze the influence of the sequence of different stages on the deraining performance, we perform experiments based on different model variants in Tab. 4. Compared to the baseline model (C-C-C), Transformer-based stage provides additional performance benefits. In addition, we also note that Transformer lacks the ability to encode local feature in the early stage, leading to suboptimal results. To ensure that CNN features and Transformer features alternate with each other in IFB, we set BaViT at the intermediate stage.



Rainy Input PReNet JORDER_E MSPFN MPRNet Uformer-B Ours

Figure 6: Visual comparison on the Internet-Data dataset. Best viewed by zooming in the figures on high-resolution displays.



Haze Input DCP DehazeNet SAR-Opt-cGAN FCTF CGAN-SAR SkyGAN Ours

Figure 7: Visual comparison on the SateHaze1k dataset. Best viewed by zooming in the figures on high-resolution displays.

C-C-C	T-C-C	C-C-T	C-T-C
39.48 / 0.974	39.62 / 0.979	39.65 / 0.981	**39.70 / 0.985**

Table 4: Ablation analysis for sequence of different stages. “C” and “T” represent CNN/Transformer-based stage.

PGB	×	✓	✓	✓	✓
CRB	×	×	✓	✓	✓
skip-connect	×	×	×	✓	✓
stage-concate	×	×	×	×	✓
PSNR (dB)	39.36	39.52	39.59	39.63	**39.70**

Table 5: Ablation analysis for different fusion components.

Effectiveness of feature fusion. To evaluate the effectiveness of our fusion strategies, we implement alternative solutions on different variants of HCT-FFN. As evident from Tab. 5, we consider the following components: (1) PGB, (2) CRB, (3) skip-connection, and (4) stage-level concatenation. Compared to direct feature concatenation (*i.e.*, w/o all these components above), our IFB (*i.e.*, PGB and CRB) tends to be more suitable for combining CNN features and Transformer features. Meanwhile, we notice that skip-connection and stage-level feature concatenation also bring out performance improvement, which shows that these operations can reduce the noise during feature propagation so that the network can adaptively learn more useful representations.

Extension to Image Dehazing

We are curious about whether our method can be extended to the image dehazing task. Here, we make comparison against different dehazing methods on the SateHaze1k (Huang et al. 2020) dataset, including DCP (He, Sun, and Tang 2010), DehazeNet (Cai et al. 2016), SAR-Opt-cGAN (Huang et al. 2020), FCTF (Li and Chen 2020), CGAN-SAR (Grohnfeldt, Schmitt, and Zhu 2018), and SkyGAN (Mehta et al. 2021). Through Tab. 6, our method produces the highest values. As shown in Fig. 7, our method can generate a clearer image.

Datasets	Thin Haze	Moderate Haze	Thick Haze
DCP	13.15 / 0.724	9.78 / 0.573	10.25 / 0.585
DehazeNet	19.75 / 0.895	18.13 / 0.855	14.33 / 0.706
Huang et al.	20.20 / 0.841	21.66 / 0.794	19.66 / 0.757
FCTF	22.77 / 0.891	24.96 / 0.932	24.14 / 0.821
CGAN-SAR	24.16 / 0.906	25.31 / 0.926	25.07 / 0.864
SkyGAN	25.38 / 0.924	25.58 / 0.903	23.43 / 0.892
Ours	**27.99 / 0.925**	**27.98 / 0.939**	**25.31 / 0.904**

Table 6: Quantitative results on SateHaze1k dataset, which contains three hazy levels, as thin, moderate, and thick haze.

Conclusion

We have presented an effective end-to-end HCT-FFN for image deraining. We introduce DaMoE modules into the CNN-based stage to emphasize the spatially-varying rain distribution, and also leverage BaViT modules into the Transformer-based stage to eliminate the spatially-long rain degradation. Importantly, a progressive IFB is involved between adjacent stages to aggregate information from CNN and Transformer. Extensive experiments demonstrate the superiority and extensibility of our method over the state-of-the-arts.

Limitation. One limitation is that the interaction between the global and local representation is explored since the hybrid fusion network only provides the final feature fusion between DaMoE and BaViT for each stage. However, the intermediate features are not considered, which are also crucial for accurate rain distribution.

Acknowledgements

This work has been supported in part by the National Key R&D Program of China (No. 2018AAA0102001), the National Natural Science Foundation of China (Nos. 61922043, 61872421), and the Fundamental Research Funds for the Central Universities (No. 30920041109).

References

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

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

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE TPAMI*, 33(12): 2341–2353.

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

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

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

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

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

Cai, B.; Xu, X.; Jia, K.; Qing, C.; and Tao, D. 2016. Dehazenet: An end-to-end system for single image haze removal. *IEEE TIP*, 25(11): 5187–5198.

Zhang, H.; and Patel, V. M. 2017. Convolutional sparse and low-rank coding-based rain streak removal. In *WACV*, 1259–1267.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep joint rain detection and removal from a single image. In *CVPR*, 1357–1366.

Fu, X.; Huang, J.; Ding, X.; Liao, Y.; and Paisley, J. 2017a. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE TIP*, 26(6): 2944–2956.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*, 7132–7141.

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

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

Li, R.; Tan, R. T.; and Cheong, L.-F. 2018. Robust optical flow in rainy scenes. In *ECCV*, 288–304.

Kim, S.; Ahn, N.; and Sohn, K.-A. 2020. Restoring Spatially-Heterogeneous Distortions Using Mixture of Experts Network. In *ACCV*.

Suganuma, M.; Liu, X.; and Okatani, T. 2019. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In *CVPR*, 9039–9048.

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

Ren, W.; Ma, L.; Zhang, J.; Pan, J.; Cao, X.; Liu, W.; and Yang, M.-H. 2018. Gated fusion network for single image dehazing. In *CVPR*, 3253–3261.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*, 7132–7141.

Wang, T.; Yang, X.; Xu, K.; Chen, S.; Zhang, Q.; and Lau, R. W. 2019. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, 12270–12279.

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

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

Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Huang, B.; Luo, Y.; Ma, J.; and Jiang, J. 2020. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 8346–8355.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Huang, B.; Zhi, L.; Yang, C.; Sun, F.; and Song, Y. 2020. Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In *WACV*, 1806–1813.

Mehta, A.; Sinha, H.; Mandal, M.; and Narang, P. 2021. Domain-aware unsupervised hyperspectral reconstruction for aerial image dehazing. In *WACV*, 413–422.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 6881–6890.

Yi, Q.; Li, J.; Dai, Q.; Fang, F.; Zhang, G.; and Zeng, T. 2021. Structure-preserving deraining with residue channel prior guidance. In *ICCV*, 4238–4247.

Quan, R.; Yu, X.; Liang, Y.; and Yang, Y. 2021. Removing raindrops and rain streaks in one go. In *CVPR*, 9147–9156.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Xiao, J.; Zhou, M.; Fu, X.; Liu, A.; and Zha, Z.-J. 2021. Improving de-raining generalization via neural reorganization. In *ICCV*, 4987–4996.

Zhou, M.; Xiao, J.; Chang, Y.; Fu, X.; Liu, A.; Pan, J.; and Zha, Z.-J. 2021. Image de-raining via continual learning. In *CVPR*, 4907–4916.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *CVPR*, 14821–14831.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *CVPRW*, 1833–1844.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *CVPR*, 12299–12310.

Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; and Wu, W. 2021. Incorporating convolution designs into visual transformers. In *ICCV*, 579–588.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *ICCV*, 22–31.

Yu, Y.; Yang, W.; Tan, Y.-P.; and Kot, A. C. 2022. Towards Robust Rain Removal Against Adversarial Attacks: A Comprehensive Benchmark Analysis and Beyond. In *CVPR*, 6013–6022.

Chen, X.; Pan, J.; Jiang, K.; Li, Y.; Huang, Y.; Kong, C.; Dai, L.; and Fan, Z. 2022. Unpaired Deep Image Deraining Using Dual Contrastive Learning. In *CVPR*, 2017–2026.

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

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image Dehazing Transformer with Transmission-Aware 3D Position Embedding. In *CVPR*, 5812–5820.

Jiang, K.; Wang, Z.; Chen, C.; Wang, Z.; Cui, L.; and Lin, C.-W. 2022. Magic ELF: Image Deraining Meets Association Learning and Transformer. In *ACM MM*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 5728–5739.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 17683–17693.

Gao, G.; Wang, Z.; Li, J.; Li, W.; Yu, Y.; and Zeng, T. 2022. Lightweight Bimodal Network for Single-Image Super-Resolution via Symmetric CNN and Recursive Transformer. *arXiv preprint arXiv:2204.13286*.

Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; and Zeng, T. 2022. Transformer for single image super-resolution. In *CVPRW*, 457–466.

Park, N.; and Kim, S. 2022. How Do Vision Transformers Work? *arXiv preprint arXiv:2202.06709*.

Huang, B.; Zhi, L.; Yang, C.; Sun, F.; and Song, Y. 2020. Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In *WACV*, 1806–1813.

Li, Y.; and Chen, X. 2020. A coarse-to-fine two-stage attentive network for haze removal of remote sensing images. *IEEE GRSL*, 18(10): 1751–1755.

Grohnfeldt, C.; Schmitt, M.; and Zhu, X. 2018. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In *IGARSS*, 1726–1729.

Gu, S.; Meng, D.; Zuo, W.; and Zhang, L. 2017. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *ICCV*, 1708–1716.

Fu, X.; Huang, J.; Zeng, D.; Huang, Y.; Ding, X.; and Paisley, J. 2017b. Removing rain from single images via a deep detail network. In *CVPR*, 3855–3863.

Yang, W.; Tan, R. T.; Feng, J.; Guo, Z.; Yan, S.; and Liu, J. 2019. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE TPAMI*, 42(6): 1377–1393.

Wang, H.; Xie, Q.; Zhao, Q.; and Meng, D. 2020. A model-driven deep neural network for single image rain removal. In *CVPR*, 3103–3112.

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