Multi-scale Residual Aggregation Deraining Network with Spatial Context-aware Pooling and Activation

Kohei Yamamichi[0000−0001−7304−6438]
a035vbu@yamaguchi-u.ac.jp
Xian-Hua Han[0000−0002−5003−3180]
hanxhua@yamaguchi-u.ac.jp

Graduate School of Science
and Technology for Innovation
University of Yamaguchi
Yamaguchi Japan

Abstract

Single image deraining is a fundamental pre-processing step in many computer vision applications for improving the visual effect and analysis performance of subsequent high-level tasks in adverse weather conditions. This study proposes a novel multi-scale residual aggregation network, to effectively solve the single image deraining problem. Specifically, we exploit a lightweight residual structure subnet with less than 10-layers as the deraining backbone network to extract fine and detailed texture context at the original scale, and leverage a multi-scale context aggregation module (MCAM) to augment the complementary semantic context for enhancing the modeling capability of the overall deraining network. The designed MCAM consists of multiple-resolution feature extraction blocks to capture diverse semantic contexts in different expanded receptive fields, and conducts progressive feature fusion between adjacent scales with residual connections, which is expected to concurrently disentangle the multi-scale structures of scene content and multiple rain layers in the rainy images, and capture high-level representative feature for reconstructing the clean image. Moreover, motivated by the fact that the adopted pooling operation and activation function in deep learning may considerably affect the prediction performance in high-level vision tasks such as image classification and object detection, we delve into a generalized pooling and activation method taking into consideration of the surrounding spatial context instead of pixel-wise operation and propose the spatial context-aware pooling (SCAP) and activation (SCAA) for incorporating with our deraining network to boost performance. Extensive experiments on the benchmark datasets demonstrate that our proposed method performs favorably against state-of-the-art (SoTA) deraining approaches.

1 Introduction

Visibility degradations arising from adverse weather such as rain, haze, and fog, significantly affect the quality of the captured images and lead to great loss of the desirable information for different computer vision applications, where the accurate surrounding context is indispensable to provide acceptable performance in real vision systems such as aerial robots, autonomous vehicles and surveillance [1]. To conquer adverse effect of the deteriorated images on the vision systems, removal of the existed rain, raindrop, or haze in the contaminated observation is a fundamental and important low-level vision task and has been extensively studied in recent years [2, 3, 4, 5, 6].

With the simple assumption of the linear mapping transformation for rainy image composite model, the observed image: **O** is generally expressed as a linear summation of the clean rain-free background: **B**, and the rain layer: **R**:

$$\mathbf{O} = \mathbf{B} + \mathbf{R} \tag{1}$$

The goal of deraining is to recover the clear image **B** from **O** via removing **R**. Since the variable number in the under-estimating components: **B** and **R** are much larger than those in the single observation, there exist infinite feasible solutions, and causes it to be a highly ill-posed problem. To restrict the solution space to valid/natural image recovery, traditional methods [4, 7, 8, 9, 10] leverage various handcrafted priors based on empirical observations to regularize the linear mapping transformation model and employ effective optimization strategy for robust image recovery. Although these prior-based methods illustrate acceptable deraining performance to some extent under controlled conditions, they usually smooth out the texture and edge details, and then cause the blurred image results. Furthermore, these methods require to conduct optimization procedure for each under-studying image, which has a high time consumption.

Recently, motivated by the great success of the deep convolutional neural network (DCNN) on image classification [11, 12, 13], object detection [14, 15, 16] and semantic segmentation [17, 18, 19], DCNN has widely applied for single image deraining as learning based paradigm [5, 8, 20, 21, 22, 23]. Benefiting from the great modeling capability and the stronger feature representation ability, the DCNN-based learning methods demonstrate remarkable performance progress for image deraining. Current effort mainly focus on designing deeper and complicated network architectures to pursue better deraining performance, and many work manifests superior deraining results with the elaborated network structure and advanced optimization (training) strategies. However, network evolution in depth and complexity unavoidably leads to substantial difficulty for practical implementation and robust model training, and also greatly increase inference time. Moreover, most current CNN models serially pile up plenty of convolutional blocks (Conv layer and activation function pairs) to learn representative features, and the increased deep stage possibly capture semantic context in large respective field. However, these deep serially connected network cannot explicitly capture the multi-scale features and context for different layers of rain, which is the latent attributes of the existed rain in observation. To handle this issue, several researchers [24, 25, 26] propose to leverage multi-scale deep framework for modeling representative features in multiple layers of rain and rich structure in the latent clear image. Unfortunately, these exploitations basically design several branches of subnets for capturing different scales of contexts, and thus result in more complicated network architectures and large model size. Moreover, current deraining work concentrates on pursuing the efficient connections among different convolutional blocks while the activation function and pooling operation [27, 28, 29, 31, 32], which have been proven to be a important aspect affecting vision task performance [33, 34, 35] , are usually un-touched and simply follow the strategies designed for high-level vision tasks. As we know that the common activations such as ReLU [27], LeakyReLU [28] and PReLU [29] usually operate on individual feature value without consideration of the surround context, and similarly the popular pooling strategy such as max pooling [31] naively take the maximum value regardless to the overall state in the target region. Although these activation and cooling functions demonstrate superior performance in the high-level vision tasks, where most existing vision systems are developed for dealing with the clear images without heavy deterioration by adverse weather, it is difficult to profess that they would be suitable for the low-level image restoration task. With the rainy images as the inputs to the deraining network, the learned features are avoidably deteriorated by some artifacts, and thus operations on the individual feature values without taking account of surround context as in conventional activation and pooling would greatly degrade the learning capability and the deraining performance.

To overcome the above limitations, this study proposes a novel multi-scale residual aggregation deraining network (MRADN) with spatial context-aware pooling and activation. We adopt a lightweight residual structure with no-so-many convolutional blocks as the deraining backbone for reducing computational cost, which focus on extracting the fine and detail texture context at the original scale. Whilst regard to the vital global semantics in the under-studying image, we exploit a multi-scale context aggregation module (MCAM) to augment the complementary semantic features for enhancing the modeling capability of the overall deraining network. Specifically, the proposed MCAM is composed in an encoder-decoder structure with multiple-resolution feature learning blocks to capture multi-scale texture and semantic contexts in diverse receptive fields. The multi-scale architecture of the MCAM is expected to concurrently disentangle the multi-scale structures of scene content and multiple rain layers in the rainy images. With the learned diverse features in both encode and decoder paths, we further conduct progressive fusion between the corresponding scale blocks of encoder-decoder paths and the adjacent-scale blocks using simple skip connections to capture high-level representative feature for reconstructing the clean image. Moreover, to suppress the potential affect of the noise and artifact on the learned feature maps with the to-be-removed rainy input, we delve in an artifact-attenuating pooling and activation method by taking consideration of the surrounding spatial context instead of pixel-wise operation, and propose the spatial context-aware pooling (SCAP) and activation (SCAA) for incorporating with our deraining network to boost performance. We conducted experiments on several benchmark datasets under different types of rains, and demonstrate the significant superiority of our method over SoTA CNN-based deraining method.

In summary, our main contributions are three-fold:

1) A novel multi-scale residual aggregation deraining network, i.e. MRADN is proposed, where the lightweight residual backbone extracts fine and detail context in the original scale while a multi-scale module learns the semantic context to complement the deficient information in the deraining backbone.

2) We design a novel multi-scale context aggregation module (MCAM) for disentangling the multi-scale structures of scene content and multiple rain layers in the rainy observation, and conduct not only the intra-module context aggregation but also integrate the aggregated multi-scale features captured in MCAM into the backbone for boosting deraining performance.

3) We propose a novel artifact-attenuating pooling and activation method via taking into account of the surrounding spatial context, dubbed as the spatial context-aware pooling (SCAP) and activation (SCAA), which is expected to be integrated into any deep learning network architecture only the available noisy input only for boosting performance.

2 Proposed multi-scale residual aggregation network

2.1 Overview

In this section, we detail the proposed multi-scale residual aggregation deraining network (MRADN). MRADN mainly consists of the residual backbone architecture, which is com-

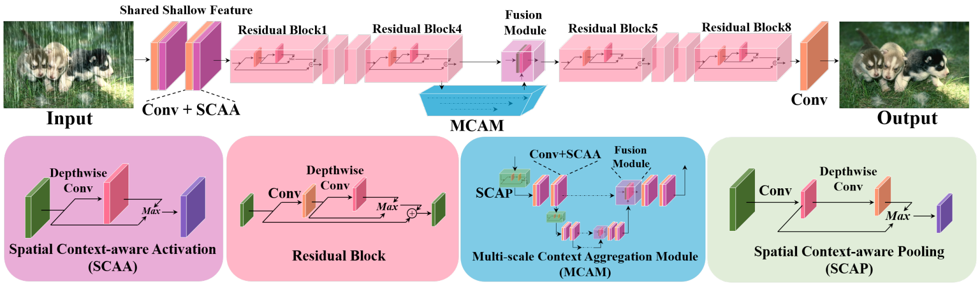

Figure 1: Conceptual architecture of our proposed MRADN.

posed with multiple residual blocks for extracting the representative features of the original resolution (0-order scale), and the multi-scale context aggregation module (MCAM), which is configured with encode-decoder structure to extract multi-scale contexts for augmenting the complementary modeling capability of the backbone. Moreover, in all blocks of the backbone and MCAM, we integrate the proposed spatial context-aware pooling (SCAP) and activation (SCAA) for emphasizing the essential features while attenuating the contaminated ones by noise or artifact instead of the conventional pooling and activation layers with point-wise operations. The conceptual structure of our proposed MRADN is shown in Fig. 1. As illustrated at the top-branch of Fig. 1, the residual backbone contains the shallow block with two convolution layers, the early-term module with four residual blocks, the late-term module with four residual blocks and a final reconstruction block with one convolution layer. In detail, the residual block consists of two convolution layers with kernel size 3×3 following the spatial context-aware activation after each, and then conducted element-wise addition with the input feature map as the output. The detail structure of the residual block is given in Fig. 1.

In overall, an input rainy image first passes through two shallow convolution layers following a SCAA function after each layer to transform the channel dimension from image to feature, and then the transformed features are inputed into early-term residual module to further extract more representative contexts at the 0-order scale of the input. Whilst the output of the early-term module is imported to the proposed MCAM with encoder-decoder structure for extracting multi-scale contexts involving information of diverse receptive fields, which is expected to disentangle the confused rain streaks and scene content in the observation, and further are progressively aggregated for being forwarded back to the residual backbone. At last, the late-term residual module and reconstruction block are seriealy adopted to transform the fused features of the early-term module and MCAM to the clear image. Next, we would describe the multi-scale context aggregation module (MCAM) and the proposed spatial context-aware pooling and activation layer for mining essential feature while attenuating noise.

2.2 Multi-scale context aggregation module

Motivated by the possible existence of the multiple decomposed rain layers especially under heavy rain conditions and aplenty of scene contents with diverse scales, the multi-scale representative context learning are preferred for reconstructing more robust clear images. As described above, the backbone network in MRADN aims to learn the detail contexts at 0-order scale (the original resolution of the input) with multiple residual blocks while cannot capture more discriminative contexts for distinguishing the rain structures and scene contents in large scale. Thus, this study investigates a multi-scale context aggregation module (MCAM) with encoder-decoder architecture to exploit and disentangle the confused rain and scene structures for extracting the discriminating and essential representation features. Specifically, the MCAM operates as a plug-and-play module to learn the complementary discriminating features with the extracted detailed contexts in the backbone network, and then automatically aggregate the essential contexts in different scales to be forwarded to the backbone again for reconstructing the clear image.

Without bells and whistles, we integrate the MCAM after the early-term module of the backbone network. In detail, given the output $\mathbf{X}^e$ of the early-term module, MCAM first adopts the spatial context-aware pooling (SCAP) to down-sample the representative features at the 0-th order scale to the first-order scale of feature $\mathbf{X}_i^1 = f_{SCAP}(\mathbf{X}^e)$, which is the input to the first-order scale of the encoder path. In overall, both encoder and decoder paths in MCAM are divided into S blocks, and each block contains 3 convolution layers with 3*3 kernels following the proposed spatial context-aware activation function after each layer. The channel number of the learned feature maps is block-wisely doubled while the spatial size is halved with the scale increasing in both encoder and decoder. The detail structure of the MCAM is shown in the light-blue background part of Fig. 2.

Let's denote the input and output of the $s-th$ order scale block in the encoder path as $\mathbf{X}_i^s$ and $\mathbf{X}_o^s$, and in the decoder as $\mathbf{Y}_i^s$ and $\mathbf{Y}_o^s$, respectively, the relation of the input and output of the $s-th$ scale block can be formulated as:

$$\mathbf{X}_o^s = F_{Conv}(\mathbf{X}_i^s, \theta_X^s), \mathbf{Y}_o^s = F_{Conv}(\mathbf{Y}_i^s, \theta_Y^s) \tag{2}$$

where $f_{Conv}(\cdot)$ represents the transformation function of 3 convolution-activation layers with the learned parameters θ_X^s and θ_Y^s, respectively. The input of the $(s+1)-th$ scale block in the encoder path is a down-sampled version from the output $\mathbf{X}_o^s$ of the $s-th$ scale block via the proposed SCAP, and is expressed as:

$$\mathbf{X}_i^{s+1} = F_{SCAP}(\mathbf{X}_o^s) \tag{3}$$

where $\mathbf{X}_i^{s+1}$ has the half size in spatial direction and double channel number of $\mathbf{X}_o^s$. Whilst the input of $s-th$ scale block in the decoder path is the fused context from the outputs of the $s-th$ scale block in the encoder path and the up-sampled output of the $(s+1)-th$ scale block in the decode path, which is formulated as:

$$\mathbf{Y}_i^s = f_{Cat}[\mathbf{X}_o^s, f_{UP}(\mathbf{Y}_o^{s+1})] \tag{4}$$

where $f_{Cat}(\cdot)$ denotes the simple concatenation operation for aggregating the feature maps in the corresponding scale of the encoder path and the previous larger scale of the decoder path while f_{UP} simply conduct bilinear up-sample operation. From Eqs. 3 to 4, the feature maps in adjacent scale blocks of the encoder path are connected using the SCAP module, and input feature maps in the decoder path are obtained by progressively aggregating the feature in the previous larger scale with the features in encoder. Thus, the final output $\mathbf{Y}_o^1$ of the first-order scale block in the decoder path would own the aggregated multi-scale contexts with diverse receptive fields. At last, $\mathbf{Y}_o^1$ is up-sampled to the resolution of the $0-order$ scale for augmenting the complementary representation of the backbone network, which is aggregated with learned feature maps $\mathbf{X}^e$ of the early-term module as:

$$\bar{\mathbf{X}}^e = f_{Cat}[\mathbf{X}^e, f_{UP}(\mathbf{Y}_o^1)] \tag{5}$$

where $\bar{\mathbf{X}}^m$ represents the input to the late-term module of the backbone network.

2.3 Spatial context-aware pooling

As introduced in section 2.2, the feature maps in the $s-order$ scale of the MCAM's encoder path are required to be spatially reduced (usually half size) for extracting representative context in larger receptive fields. The generic method for decreasing spatial size in most high-level vision tasks such as image classification and object detection popularly adopt the average or max pooling layer and their variants, and verify impressive performance in different applications with the clear inputs. The conventional pooling methods simply conduct comparison on multiple values of a local spatial region without taking into account of the possible contamination by noise. However, in our under-study deraining scenario, the inputs are the rainy images, and thus may lead to the learned feature maps in the network to be polluted by worthless artifact.

This study aims to exploit a spatial context-aware pooling strategy to decrease influence of the noise on the spatial size reduced maps. Specifically, with a feature map $\mathbf{X} \in \Re^{W \times H \times C}$, we want to aggregate multiple features in a spatial local region to produce a more compact representation. The widely used max pooling layer simply takes the maximum value to capture most salient activation such as giving one maximum of 4 activations in a 2×2 local region. Although the compact feature reflects the active status of a local region, it may be activated by the unwanted interference such as the rain streak in our deraining scenario. Instead of conducting max operation by comparison of the individual values, we firstly adopt a convolution layer with kernel size 2×2 and stride 2, which aggregates the features of the local region with a learnable weights into one representative activation, and produce a compact representation $\hat{\mathbf{X}} \in \Re^{W/2 \times H/2 \times C}$ as follows:

$$\hat{\mathbf{X}} = f_{k2s2}(\mathbf{X}) \tag{6}$$

Moreover, to integrate more spatial contexts in a larger scale, we further exploit a depth-wise convolution layer with kernel size 3×3 on $\hat{\mathbf{X}}$, and then conduct element-wise max operation on $\hat{\mathbf{X}}$ and its depth-wise convoluted feature maps, which is formulated as follows:
.

$$\tilde{\mathbf{X}} = max(\hat{\mathbf{X}}, f_{DW}(\hat{\mathbf{X}})) \tag{7}$$

where $\tilde{\mathbf{X}} \in \Re^{W/2 \times H/2 \times C}$ is the resulted compact feature maps with our proposed SCAP. The conceptual structure of the SCAP is shown in the light-green background part of Fig. 1.

2.4 Spatial context-aware activation: SCAA

Activation function is an essential component in the modern CNN network, and our deraining network also employs activation layers after all convolution layers. Thus, an effective activation function for the low-level vision task with the noisy inputs such as confused rain streak in the inputs would be critical aspect to affect deraining performance. Given the feature map $\mathbf{X}^c$ extracted by a convolution layer, the widely used ReLU activation simply maintains the positive features and sets all negative features as zero, which cannot fully explore the surrounding spatial context to produce a robust activation, and is mathematically expressed as:

$$\mathbf{X}^{ReLU} = max(\mathbf{X}^c, \mathbf{0}) \tag{8}$$

where $\mathbf{0}$ denotes a matrix of the same size with $\mathbf{X}^c$, and all elements are 0. To incorporate the surrounding context into consideration, our SCAA firstly adopts a depth-wise convolution on

$\mathbf{X}^c$ to produce spatial context aggregated feature map: $f_{DW}(\mathbf{X}^c)$, and then conduct element-wise max operation on $\mathbf{X}^c$ and $f_{DW}(\mathbf{X}^c)$, as shown in the magenta background part of Fig. 1. The formula of the SCAA is expressed as

$$\mathbf{X}^{SCAA} = max(\mathbf{X}^c, f_{DW}(\mathbf{X}^c)) \tag{9}$$

We integrate the proposed SCAA in all convolution blocks of our deraining network. Since only one depth-wise convolution layer is additionally acquired, the parameter increasing due to the SCAA can be neglected compared with the overall network's parameters.

3 Experimental Results

In this section, we will conduct extensive experiments to demonstrate the effectiveness of our proposed multi-scale residual aggregation deraining network. we first introduce the experimental setting including the used datasets, evaluation metrics and detail implementation, and then provide the comparisons with the state-of-the-art deraining methods and ablation study.

Table 1: Average PSNR and SSIM comparison on the synthetic datasets Rain1200, Rain200L, Rain200H, and Rain800. Red and blue colors are used to indicate top 1^{st}, 2^{nd} performance.

Methods	Rain1200		Rain200L		Rain200H		Rain800		#.Parameters
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
Rainy	23.64	0.727	26.71	0.834	13.79	0.367	22.18	0.663	-
DDN[5]	23.39	0.832	34.46	0.957	26.11	0.792	22.78	0.803	58,175
DIDMDN[36]	29.66	0.899	35.40	0.962	26.61	0.824	22.53	0.812	135,800
RESCAN[41]	30.54	0.879	36.09	0.970	26.75	0.835	24.99	0.830	499,668
UMRL[42]	30.57	0.909	33.83	0.957	23.01	0.744	24.99	0.869	984,356
PReNet[43]	31.49	0.910	37.25	0.978	28.57	0.887	24.79	0.849	168,963
SPANet[44]	31.94	0.902	35.79	0.965	26.27	0.865	22.41	0.838	283,716
MSPFN[25]	32.06	0.913	31.64	0.925	27.39	0.843	27.01	0.851	15,823,424
MPRNet [45]	32.94	0.914	35.72	0.962	29.49	0.887	29.61	0.874	3,637,249
MCGKT-Net[46]	32.91	0.916	37.13	0.973	28.71	0.873	28.73	0.868	14,761,155
MRADN	34.33	0.931	39.44	0.985	29.69	0.900	29.66	**0.897**	8,306,051
LW-MRADN (Common)	32.65	0.919	36.13	0.974	28.65	0.871	28.97	0.879	3,420,547
LW-MRADN	**34.76**	**0.937**	**39.54**	**0.986**	**29.92**	**0.904**	**29.70**	0.896	3,420,547

3.1 Experimental setting

Datasets: We carry out experiments on three deraining datasets: Rain1200 [36], Rain200L [37], Rain200H [37], and Rain800 [38]. Rain1200 dataset includes 12000 images for training and 1200 images for testing, and the rainy images are generated with different levels of rainy density under light, medium and heavy rain conditions. The images in Rain200L has light rain and is relatively easy dataset. The training subset contains 1800 image pairs and the test subset has 200 images. Rain 200H has the same number of training and testing images but being contaminated by more heavy rain with different shapes, directions, and sizes, and thus is the most challenging dataset in deraining community. Rain800 consist of in total 800 images with 700 rainy/clean pairs as the training samples and the remainders as testing. The rainy images in Rain800 are synthesized by adding fine rain streak to the clean images following the guidelines mentioned in [20], and have the fine-grained streaks with noise-like structures.

Evaluation metrics: We adopt two commonly used evaluation metrics: i.e. peak signal to noise ratio (PSNR) and structure similarity index (SSIM [39]) to assess the performance

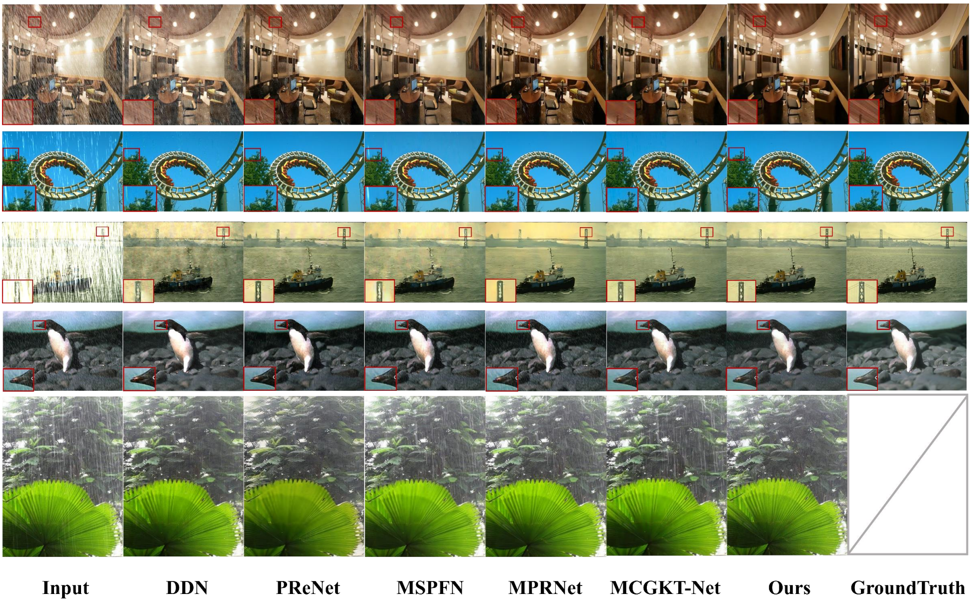

Figure 2: Qualitative comparison with the state-of-the-art methods. The first to forth rows are the results of four images from the synthetic datasets while the fifth row manifests the recovered results of a real rainy image [6] without the corresponding ground-truth image.

of our deraining method quantitatively. SSIM evaluates the image structure difference and is more consistent with human perceptual measure. Note that as the human visual system is sensitive to the Y channel of a color image in YCbCr space, we calculate PSNR and SSIM with the converted luminance (Y) channel only.

Training details: We use Keras with TensorFlow backbend to train and test our proposed method. In the training process, we crop 256×256 patches from the training samples, and adopt Adam [40] to optimize our network. The networks are trained with 500 epochs and the learning rate is set as 2×10^{-4}. The MAE between the network reconstructions and the ground-truth clear images is used as the loss function for network training.

3.2 Evaluation comparisons with state-of-the-art

We compare our proposed MRADN with the state-of-the-art methods, including deep detail network (DDN) [6], density-aware deraining (DIDMDN) [36], recurrent squeeze-and-excitation context aggregation net (RESCAN) [41], progressive deraining network (PreNet) [43], spatial attentive network (SPANet) [44], multi-scale progressive fusion network (MSPFN) [25], multi-stage progressive restoration network (MPRNet) [45], and multi-level context gating knowledge transfer network (MCGKT-Net) [46].

The deraining models are separately trained with the training pairs in the datasets: Rain200H, Rain200L, Rain800 and Rain1200, and the quantitative results are calculated with the learned model under the corresponding dataset, respectively. The quantitative metrics of our and the compared deraining methods are manifested in Table 1. It is obvious that our proposed MRADN has illustrated the highest SSIM and PSNR in all datasets. Compared with the state-of-the-art methods, our approach achieves a great improvement over most methods. In

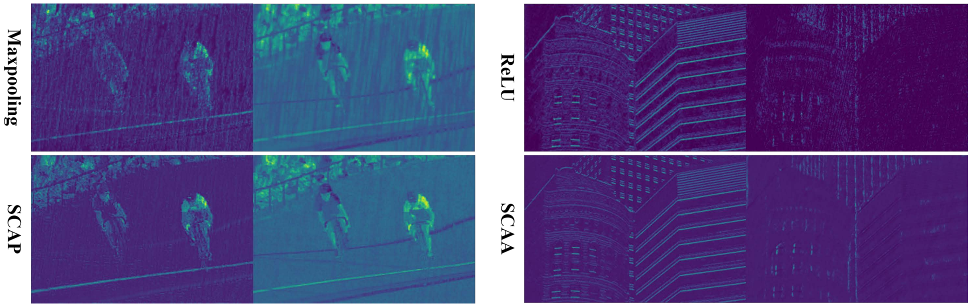

Figure 3: The compared feature maps using the conventional pooling/activation and the proposed SCAA/SCAP.

addition, it can observe that our proposed model needs more parameters than some SoTA methods, which are resulted from MCAM according to our analysis. Thus, to decrease the model parameters, we replace the vanilla convolution layers in the MCAM using the GhostConv layer [47], and exploited a lightweight MCAM to form our LW-MRADN. The compared performance and model parameter are also illustrated in Table 1, which demonstrates that LW-MRADN can achieve better performance than the model with similar size such as (MPRNet) [45]. In addition, to produce a common model with high generalization capability, we further combined different types of rain image datasets to train our LW-MRADN, named as LW-MRADN (Common), and then recover the clear images for all four datasets. The quantitative performance of all four datasets using the common model are also provided in Table 1. Further, the common model is also adapted for recover the clear image for a real dataset (15 images) [6]. The visualization examples with our network and different SoTA methods have been shown in Fig. 2 on both synthetic images and a real image. From Fig. 2, we can see that the proposed model can restore clearer results.

Table 2: Ablation study w/o the SCAP and SCAA.

SCAA		×	✓	×	✓	LReLU	PReLU	✓	✓
SCAP		×	×	✓	✓	✓	✓	AP	WP
Rain200H	PSNR	29.02	29.58	29.46	**29.69**	29.45	29.33	29.54	29.59
	SSIM	0.891	**0.901**	0.895	0.900	0.896	0.897	0.900	0.901
Rain200L	PSNR	38.73	39.11	38.38	**39.44**	39.28	39.26	39.14	39.34
	SSIM	0.983	0.985	0.982	**0.985**	0.984	0.985	0.984	0.985
#. Parameters(M)		8.00	8.03	8.27	8.31	8.27	8.27	8.03	9.51

3.3 Ablation Studies

In this section, we evaluate the effectiveness of different proposed modules. Especially, we utilize the simple network consisting of multiple residual blocks as the baseline module (the top branch in Fig. 1), and conduct two kinds of ablation studies for verifying the contributions of the proposed SCAP and SCAA, and the MCAM with different scales ($s = 1,2,3$). Basically, we consider the max pooling (MP) and the ReLU activation as the default method, and further conducted other pooling operations such as average pooling (AP) and Wavelet-based pooling (WP) [48] as well as other activation function such LReLU(LeakyReLU) and PReLU for comparison. Table. 2 manifests the compared quantitative values of our MRADM with the proposed SCAP/SCAA or the conventional activations (ReLU/LReLU/PReLU) and the pooling layers (MP/AP/WP) for feature map down-

Table 3: Ablation results integrating the SCAA and SCAP with different CNN models on the Rain200H dataset.

(a) Ablation results integrating the SCAA

Methods	Default (ReLU)		SCAA	
	PSNR	SSIM	PSNR	SSIM
DDN[5]	26.11	0.792	**27.13**	**0.796**
UMRL[42]	23.01	0.744	**24.25**	**0.752**
MCGKT-Net[46]	28.71	0.873	**29.08**	**0.894**

(b) Ablation results integrating the SCAP

Methods	Default (MaxPooling)		SCAP	
	PSNR	SSIM	PSNR	SSIM
UMRL[42]	23.01	0.744	**23.95**	**0.749**
MCGKT-Net[46]	28.71	0.873	**28.88**	**0.889**

Table 4: Ablation study w/o the MCAM and different scales in the MCAM.

Scale	Baseline(s=0)	MCAM (s=1)	MCAM (s=2)	MCAM (s=3)
PSNR	28.12	29.01	29.47	**29.69**
SSIM	0.870	0.888	0.897	**0.900**
#. Parameters(M)	0.34	0.72	2.26	8.31

sampling in MCAM module, which shows that the integration of the SCAP and SCAA instead of the conventional methods do not greatly increase the network parameters while stably improve the restoration results with almost 1dB PSNR for Rain200L dataset. Table. 3 demonstrates the compared results by integrating the SCAA/SCAP with different CNN models. Moreover, Fig. 3 provides some learned feature maps using the conventional max-pooling, ReLU activation and our proposed SCAA and SCAP layers, which in turn manifest our proposed SCAA/SCAP can alleviate the noise and artifact. Table. 4 provides the compared result w/o MCAM modul and different scales in the MCAM module, and validate that the aggregation with large scale in MCAM is capable of gradually boosting the restoration performance. Compared with the baseline module, the introduced MCAM improve the PSNR about 1.57 and SSIM about 0.03.

4 Conclusion

In this work, we proposed a novel multi-scale residual aggregation network, to effectively solve the image deraining problem. Specifically, we exploited a residual subnet with a few blocks as the backbone architecture and a multi-scale context aggregation module (MCAM) to augment complementary semantic context for enhancing capability of the network. Futhermore, we delved in a generalized pooling and activation method taking consideration of the surrounding spatial context instead of pixel-wise operation, and propose the spatial context-aware pooling (SCAP) and activation (SCAA) for incorporating with our deraining network to boost performance. Extensive experiments on the benchmark datasets demonstrated that our proposed method performs favorably against state-of-the-art deraining approaches.

Acknowledge

This research was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20K11867, and JSPS KAKENHI Grant Number JP12345678.

References

[1] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detectionin surveillance videos. In CVPR, 2018.

[2] Y.-L. Chen and C.-T. Hsu. A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In ICCV, 2013.

[3] Y. Luo, Y. Xu, and H. Ji. Removing rain from a single image via discriminative sparse coding. In ICCV, 2015.

[4] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown. Rain streak removal using layer priors. In CVPR, 2016.

[5] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley. Removing rain from single images via a deep detail network. In CVPR, 2017.

[6] 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 CVPR, 2017.

[7] L. W. Kang, C. W. Lin, and Y. H. Fu. Automatic single image-based rain streaks removal via image decomposition. IEEE Transactions on Image Process-ing, 21(4):1742–1755, 2012.

[8] D. A. Huang, L. W. Kang, Y. C. F. Wang, and C. W. Lin. Self-learning based image decomposition with applications to single image denoising. IEEE Trans. Multimedia16(1): 83–93, 2014.

[9] Y. Chang, L. Yan, and S. Zhong. Transformed low-rank model for line pattern noise removal. In ICCV, 2017.

[10] L. Zhu, C.-W. Fu, D. Lischinski, and P.-A. Heng. Joint bi-Layer optimization for single-image rain streak removal. In ICCV, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NeurIPS, 2012.

[12] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In NeurIPS, 2018.

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In arXiv preprint arXiv:1709.01507, 2017.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR 2014.

[15] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In CVPR, 2018.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In ECCV, 2018.

[17] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In CVPR, 2019.

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. CCNet: Criss-cross attention for semantic segmentation. In ICCV, 2019.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[20] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley. Clearing the skies: A deepnetwork architecture for single-image rain removal. IEEE Transactions on ImageProcessing, pp. 2944–2956, 2017.

[21] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha. Recurrent Squeeze-and-Excitation Context Aggregation Net for Single Image Deraining. In ECCV, 2018.

[22] 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 ACMMM, 2018.

[23] W. Yu, Z. Huang, W. Zhang, L. Feng, and N. Xiao. Gradual network for single image de-raining. In ACMMM, 2019.

[24] Y. Zheng, X. Yu, M. Liu, and S. Zhang. Residual multiscale basedsingle image deraining. In BMVC, 2019.

[25] K. Jiang, Z. Wang, P. Yi, B. Huang, Y. Luo, J. Ma, and J. Jiang. Multi-scale progressive fusion network for single image deraining. In CVPR, 2020.

[26] X. Fu, B. Liang, Y. Huang, X. Ding, and J. Paisley. Lightweight pyramid networks for image deraining. In TNNLS, 2019.

[27] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, 2010.

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In ICML, 2013.

[29] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In ICCV, 2015.

[30] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature, 405(6789):947–951, 2000.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for ImageRecognition. In CVPR, 2016.

[32] N. Murray, H. Jegou, F. Perronnin, and A. Zisserman. Interferences in Match Kernels. PAMI, vol. 39, no. 9, pp. 1797–1810, 2016.

[33] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. In arXiv preprint arXiv:1710.05941, 2017.

[34] D. A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In ICLR, 2016.

[35] N. Ma, X. Zhang, M. Liu, J. Sun. Activate or not: Learning customized activation. In CVPR, 2021.

[36] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. In CVPR, 2018.

[37] 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 CVPR, 2017.

[38] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. In arXiv preprint arXiv:1701.05957, 2017.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Trans. on ImageProcessing, vol. 13, no. 4, pp. 600–612, April 2004.

[40] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[41] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In ECCV, 2018.

[42] R. Yasarla, and V. M. Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In CVPR, 2019.

[43] D. Ren, W. Zuo, Q. Hu, P. Zhu, D. Meng. Progressive image deraining networks: a better and simpler baseline. In CVPR, 2019.

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

[45] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Multi-stage progressive image restoration. In CVPR, 2021.

[46] K. Yamamichi, and X.-H. Han. Multi-level context gating knowledge transfer network for single image deraining. In ACCV, 2020.

[47] K. Han, Y. Wang, Q. Tian, J. Guo, Chunjing. Xu, and Chang. Xu. GohstNet: More Features from Cheap Operations. In CVPR 2020.

[48] P. Liu, H. Zhang, W. Lian, and W. Zuo. Multi-level wavelet convolutional neural networks. In CVPR Workshops, 2018.