Lightweight Multi-Scale Context Aggregation Deraining Network With Artifact-Attenuating Pooling and Activation Functions

KOHEI YAMAMICHI AND XIAN-HUA HAN, (Member, IEEE)
Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi 753-8511, Japan
Corresponding author: Xian-Hua Han (hanxhua@yamaguchi-u.ac.jp)

This work was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture, and Sports (MEXT) under Grant 20K11867, and in part by JSPS KAKENHI under Grant JP12345678.

ABSTRACT Single image deraining is a fundamental pre-processing step in many computer vision applications for improving the visual effect and system performance of the downstream high-level tasks in adverse weather conditions. This study proposes a novel multi-scale context aggregation network, to effectively solve the single image deraining problem. Specifically, we exploit a lightweight residual structure subnet as the baseline architecture to extract fine and detailed texture context at the original scale and further incorporate a multi-scale progressive aggregation module (MPAM) to learn the complementary high-level context for enhancing the modeling capability of the overall deraining network. The MPAM, designed as a plug-and-play module to be utilized in the arbitrary network, is composed of multi-scale convolution blocks to learn a wide variety of contexts in multiple receptive fields, and then carries out progressive context aggregation between adjacent scales with residual connections, which is expected to concurrently disentangle the multi-scale structures of scene contents and multiple rain layers in the rainy images, and models more representative contexts for reconstructing the clean image. To reduce the learnable parameters in the MPAM, we further explore a context hallucinate block for replacing the multi-scale convolution block, and propose a lightweight MPAM. Moreover, for being specially adaptive to deal with the input rainy images with a lot of unwanted components (rain layers), we delve into the artifact-attenuating pooling and activation functions via 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 deraining approaches.

INDEX TERMS Deep residual block, multi-scale progressive aggregation, context hallucinate block, artifact-attenuating pooling and activation, image deraining.

I. INTRODUCTION

Visibility degradations arising from adverse weather such as rain, haze, and fog, extremely affect the quality of the captured images and cause great loss of the desirable information for various computer vision applications, where the accurate surrounding context is indispensable to provide acceptable performance in real vision systems such as autonomous vehicles, surveillance, and aerial robots [1], [2]. To defeat the adverse effect of the degraded images on the high-level vision systems, reduction of the existed rain, raindrop, or haze in the contaminated observation is a fundamental low-level vision task and has extracted remarkable research attention in recent years [3]–[5], [8], [12].

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: $R$:

$$O = B + R$$

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], [10], [11], [13], [14] 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 procedures 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 [15], [16], [42], object detection [18]–[20] and semantic segmentation [21]–[23], DCNN has widely applied for single image deraining as learning-based paradigm [5], [7], [11], [12], [24], [26], [27]. 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 efforts mainly focus on designing deeper and complicated network architectures to pursue better deraining performance, and many works manifest superior deraining results with the elaborated network structures 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 increases inference time. Moreover, most current CNN models serially pile up plenty of convolutional blocks (Conv layer and activation function pair) to learn representative features, and the increased deep stages possibly capture the semantic context in the large respective fields. However, these deep serially connected networks cannot explicitly capture the multi-scale features and contexts for different layers of rain, which are the latent attributes of the existed rain in observation. To handle this issue, several researchers [28]–[30] proposed to leverage multi-scale deep architectures for learning representative features in multiple layers of rain and rich contexts in the latent clear image. Unfortunately, these methods generally utilize several branches of subnets to model different scales of contexts and thus lead to more complicated network architectures and high computational costs. In addition, the existing deraining networks concentrate on pursuing the efficient connections among different convolutional blocks while the activation function and pooling operation [31]–[35], which have been proven to be an important aspect affecting vision task performance [36]–[38], are usually untouched and simply follow the strategies designed for high-level vision tasks. As we know that the popularly used activation functions such as ReLU [31], LeakyReLU [32], and PReLU [33] basically operate on individual feature values regardless of the feature response of the surrounding context, and similarly, the existing pooling strategy i.e. max pooling [34] naively adopts the maximum value without taking account of the overall state in the target region. Although these activation and pooling functions manifest promising performance in the high-level vision tasks, which are mainly used for dealing with clear images without heavy deterioration by adverse weather, it is difficult to profess that they would be adaptable to the low-level image restoration task. With the rainy images as the inputs to the deraining network, the learned features are unavoidably contaminated by some artifacts, and thus operations on the individual feature values without considering the surrounding context as in conventional activation and pooling would possibly destroy the learning capability of effective features and further degrade the deraining performance.

To solve the above issues, this study proposes a novel lightweight multi-scale context aggregation deraining network (MCADN) with artifact-attenuating pooling and activation functions. Specifically, we employ a lightweight residual structure with not-so-many convolution blocks as the baseline architecture to focus on extracting the fine and detailed texture context at the original scale, and then exploit a multi-scale progressive aggregation module (MPAM) to capture the complementary high-level (semantic) features for enhancing the modeling capability of the overall deraining network. The proposed MPAM is feasibly used as a plug-and-play module in the whole deraining network and is architectured as an encoder-decoder layout with multi-scale convolution blocks to learn a wide variety of textures and contexts in diverse receptive fields. With its inherent multi-scale attributes, MPAM has the prospected advantage to concurrently disentangle the multi-scale structures of scene contents and multiple rain layers in the rainy images, and then produces the comprehensive contexts to be feed-forwarded to the backbone residual subnet via progressive fusion between the corresponding scale blocks of encoder-decoder paths and the adjacent-scale blocks. In addition, to reduce the parameter numbers in the multi-scale architecture, we propose a lightweight MPAM by exploring a context hallucinate block instead of the vanilla convolution block. Moreover, to suppress the potential effect of the noise and artifact on the learned feature maps with the to-be-removed rainy input, we delve into the artifact-attenuating pooling and activation functions by 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. We conduct extensive experiments on several benchmark datasets under different types of rains and demonstrate the significant superiority of our method over SoTA CNN-based deraining methods. Fig. 1 provides the compared deraining performances and model sizes of our method and the state-of-the-art (SoTA) deep learning models.

In summary, our main contributions are four-fold:
1) A novel lightweight multi-scale context aggregation deraining network, i.e. MCADN is proposed, which aims to extract fine and detailed context in the original scale with a
lightweight residual backbone while capture the complementary semantic context with a multi-scale module.

2) We leverage a novel multi-scale progressive aggregation module (MPAM) for disentangling the multi-scale structures of scene contents and multiple rain layers in the rainy observation, and conduct not only the intra-module context aggregation but also feed-forward the aggregated multi-scale features in MPAM to the backbone subnet for boosting deraining performance.

3) We exploit a context hallucinate block instead of the vanilla convolution blocks in the MPAM, and propose a lightweight MPAM to reduce the heavy parameter load.

4) 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 can be integrated into arbitrary deep network architecture with the available noisy input only for performance boosting.

II. RELATED WORK

Single image deraining aims to recover the clear (rain-free) images from the rainy observation images, and has witnessed significant evolution in recent years. This section briefly surveys the single image deraining paradigms, multi-scale learning networks, and the existed pooling and activation operations.

A. SINGLE IMAGE DERAINING

Removal of rain streaks or drops from a single image is a great challenging task due to its ill-posed nature. There have proposed numerous approaches to tackle this problem, which can mainly be categorized into prior-based methods [4], [10]–[13] and learning-based methods [5]–[9]. Prior-based image deraining methods usually are to optimize the physical imaging model, which is formulated according to the fact that rainy images can be decomposed into a clean background image layer and a rain layer, and leverage various handcraft priors from empirical observation for recovering the robust clear images. Kim et al. [39] utilized a simple adaptive nonlocal mean filter for single image deraining. Most methods attempt to separate rain streak layers from rainy images via modeling the underlying structure of the latent clean image and the attribute of the rain layers. For example, Kang et al. [10] proposed a dictionary learning-based morphological component analysis to remove rain streaks in high-frequency regions while Huang et al. [11] investigated a self-learning-based image decomposition method to distinguish rain streaks. Luo et al. [12] proposed a discriminative sparse coding for modeling image patches to recognize rain streaks from non-rain content. Chen and Hsu [3] and Chang et al. [13] leveraged the low-rank properties to model and separate rain streaks. Further, to accommodate multiple orientations and scales of rain streaks, Li et al. [4] introduced a GMM to model layer prior while Wang et al. [40] proposed a dictionary learning-based hierarchical scheme to progressively remove rain and snow. Later, Zhu et al. [14] integrated three priors to a joint optimization process for single image rain removal. Although the physical imaging models [4], [10]–[14] regularized by the empirical priors such as the sparseness, self-similarity, and low-rankness demonstrate recovering performance improvement with the elaborated priors, it is still insufficient in modeling the rich properties or structures of the decomposed layers in the diverse images. Moreover, the explored priors are usually hand-crafted, and to hammer our a appreciate prior for a specific image, and a kind of rain type still remains to be an art.

With the prosperity of the deep convolutional neural network (DCNN) in different computer vision applications, DCNN has also widely applied for single image deraining, and been validated remarkable performance progress [11], [12], [24]. Fu et al. [24] first conducted a pilot study, which adopted a simple three-layer convolutional network on high-frequency parts to predict clean components from its rain-contaminated counterpart, and further extended it to combine deep residual-block and a global skip connection [5] for boosting deraining performance. The current deraining researches mainly focus on designing advanced and complicated network architecture. Yang et al. [8], [41] proposed a recurrent dilated network with multi-task learning to jointly detect and remove the rain from a single image, and Li et al. [7] combined the recurrent neural network architecture with squeeze-and-excitation block [42] for rain removal. Zhang and Patel [6] introduced a multi-stream dense network for joint rain density estimation and deraining. To improve the visual quality of the derained images, the generative adversarial network [43] was incorporated to integrate the discriminative capability into the loss function for pursuing more sharp and natural images. In [44], a residual-guide network with recursive modules and multi-level supervision has been proposed to progressively predict high-quality derained images while a non-locally enhanced encoder-decoder network [26] was investigated to capture global spatial dependencies [27] for accurate rain streaks estimation. Wei et al. [45] exploited a
semi-supervised image deraining paradigm to handle the generalization issue on the degraded performance of the learning methods for unseen rain types and Ren et al. [46] investigated several factors such as different network architectures: progressive ResNet and recurrent blocks inside and cross stages, and loss functions for probing a better and simpler baseline deraining network. More recently, Yang et al. [47], [48] proposed a fractal band learning network with self-supervision to extract scale-robust rain features whilst Yasarla et al. [49] introduced a semi-supervised learning deraining method based on the Gaussian process. Later, Wang et al. [65] presented a model-based interpretable network via integrating the conventional convolutional dictionary-based methods and deep learning, and Fu et al. [50] propose a dual graph convolutional network via designing two orthogonal graphs to perform global relational modeling and reasoning for rain streak removal. In addition, to enhance the generalization capability of the learned CNN model, Gan [25] leveraged meta-learning-based relation and representation learning networks for single-image deraining, and demonstrated superior deraining performance.

B. MULTI-SCALE LEARNING NETWORK

Inspired by the observation that a rainy image is possibly decomposed into the image layer with different types of contents and multiple layers of rain streaks, explicitly modeling the separable components according to their corresponding characteristics would benefit to extract more interpretable and representative features, and thus boost the deraining performance. To handle this issue, several researchers have investigated multi-stage image deraining networks in a progressive manner [28]–[30], [46], [51]. However, these methods usually adopt multiple similar subnets (several main-streams) to concentrate on extracting different scales of contexts and unavoidably result in more complicated DCNN architectures. Moreover, the adopted identical subnets in different stages cannot explicitly take into account of the specific attributes of different composed components such as the detailed or salience contexts with various receptive fields, and thus may yield suboptimal results. It is well known that the widely used encoder-decoder architectures such as U-Net [23] and FCN [52] in semantic image segmentation have naturally multi-scale learning capability, and can explicitly extract the representative features in diverse resolution contexts from the original input resolution to decimal decreased scales. To unitize the composing module in each scale, the encoder-decoder architecture generally adapts identical convolutional block in different scales and may lead to over-simplified sub-block to provide sufficient representative features in the original input resolution, which is critically important to recover the detailed texture of the target in low-level vision applications such as our deraining task. To effectively leverage the advantage in the serial connected CNN and the encoder-decoder architecture, this study exploits a lightweight residual structure subnet as the deraining backbone to extract fine and detailed texture context at the original scale, and leverages an encoder-decoder-like architecture as the multi-scale progressive aggregation module (MPAM) to augment the complementary semantic context for enhancing the modeling capability of the overall deraining network.

C. POOLING AND ACTIVATION FUNCTIONS

Pooling and activation operations are the indispensable components in the modern deep CNN models. Pooling layers are used for aggregating activation of spatial locations to produce more compact feature maps, and mainly include average [53] and max pooling [34]. Some works attempt to learn a pooling function from the data using the neural network optimization framework. For example, Lee et al. [54] presented to learn a tree-like pooling filter to conduct the spatial pooling while Christlein et al. [55] exploited a generalized max pooling method via learning weights in each local region through an optimization process [35]. Moreover, Pinheiro and Collobert [56] introduced a log-sum-exp (LSE) pooling to compute a smooth approximation of the max function. All these pooling strategies are developed for boosting the performance of the high-level vision tasks with clear data as input.

On the other hand, regarding to activation functions, the Rectified Linear Unit (ReLU) [32], [57], [58] and its variants [33], [37], [59] are the most popular operations in the past few years. Based on the LogSumExp function, many methods [60]–[63] proposed to smoothly approximate the maximum function ReLU. Goodfellow et al. [64] introduced a piecewise linear approximation for arbitrary convex activation functions while Ramachandran et al. [36] adopted the recent architecture research techniques to figure out a novel scalar activation called Swish. Further, Ma et al. [38] generalized Swish [36] to the ACON family, and proved the generalized activation is a smooth approximation of ReLU based on the well-known smooth conversion called $\alpha - softmax$. Similar to pooling function researches, the effectiveness of all the proposed activation functions are evaluated on high-level vision tasks with clear inputs without heavy deterioration by adverse weather, and usually operated on the individual feature values without taking into account of surrounding contexts. Our goal in this study is to recover a clear image from a heavy deteriorated input, which may lead to noisy values or artifacts even in the learned feature maps. Therefore, it is hard to profess that the widely used pooling and activation layers would be suitable for the low-level image restoration task. This study aims to exploit a Spatial Context-aware Pooling and Activation via incorporating the surrounding spatial contexts for attenuating the existing noise and artifact.

III. PROPOSED MULTI-SCALE CONTEXT AGGREGATION NETWORK

A. OVERVIEW

In this section, we present in detail the proposed multi-scale context aggregation deraining network (MCADN). MCADN mainly consists of two parts: the residual backbone subnet,
which is composed of multiple residual blocks for learning the representative features at the original resolution (0-order scale), and the multi-scale progressive aggregation module (MPAM), which is structured with encode-decoder architecture to learn multi-scale contexts for enhancing the complementary modeling capability of the backbone. In all blocks of the backbone subnet and MPAM, we incorporate the artifact-attenuating pooling and activation functions: the spatial context-aware pooling (SCAP) and activation (SCAA) for emphasizing the useful and important features while attenuating the contaminated ones by noise or artifact instead of the conventional pooling and activation layers with point-wise operations. Moreover, we deploy the proposed context hallucinate block for replacing the vanilla convolution blocks in MPAM to reduce the heavy parameter load. The conceptual framework of our proposed MCADN is illustrated in Fig. 2. As shown in the top-branch of Fig. 2, the residual backbone subnet includes a shallow module with two convolution layers, an early-phase module with four residual blocks, an optional middle-phase module with three residual blocks, a late-phase module with four residual blocks, and a final reconstruction module with one convolution layer. In detail, the residual block is composed of one vanilla convolution layer with kernel size $3 \times 3$ following the spatial context-aware activation function, and then the output feature is calculated via element-wise addition with the input feature map. The detailed structure of the residual block is illustrated as the second part of the bottom branch in Fig. 2.

Overall, an input rainy image firstly goes through the shallow module, where a SCAA is utilized after each convolution layer, to transform the RGB channel from image to features, and then the transformed features are inputted into the early-phase module to further extract more representative contexts at the 0-order scale of the input. The output of the early-phase module is imported to the optional middle-phase module and the proposed MPAM for learning multi-scale contexts involving information of diverse receptive fields, which is expected to disentangle the confused rain layers and scene contents in the observation, and then the progressively aggregated multi-scale context within the MPAP is forwarded back to the backbone subnet. Finally, the late-phase residual module and reconstruction block are serially adopted to transform the fused features of the previous module and MPAM to the clear image. In the following subsection, we would present the multi-scale progressive aggregation module (MPAM) and the proposed artifact-attenuating pooling and activation layers for mining more effective and important features.

### B. MULTI-SCALE PROGRESSIVE AGGREGATION MODULE

With the possible decomposition of multiple rain layers especially under heavy rain conditions and aplenty of scene contents with diverse scales, disentangling different components of the rainy images is preferred for reconstructing more robust clear images, and multi-scale representative context learning in deep networks has the potential to handle this issue. As introduced above, the backbone subnet in our MCADN aims to learn the detail contexts at 0-order scale (the original resolution of the input) with multiple residual blocks but cannot explicitly separate the confusing multi-scale contexts of the rain structures and scene contents. In this study, we incorporate a multi-scale progressive aggregation module (MPAM) with encoder-decoder architecture to exploit and disentangle the confused rain and scene structures, and then progressively aggregate the essential and effective features as the comprehensive contexts to be forwarded to the backbone subnet. The MPAM operates as a plug-and-play module to learn the complementary multi-scale context of the extracted detailed contexts in the backbone network and thus can enhance the modeling capability for recovering more accurate images. Without bells and whistles, we employ the basic MPAM after the early-phase module of the backbone subnet. Next, we present the detailed architecture of the basic MPAM and an improved lightweight version (LW-MPAM) to reduce the learnable parameters.
1) The Basic MPAM

In detail, given the output feature \( X^e \) of the early-phase module, MPAM firstly use the spatial context-aware pooling (SCAP) to down-sample \( X^e \) at the 0-th order scale as \( X_1^e = f_{SCAP}(X^e) \), which is the input to the first-order scale convolution block of the encoder. Overall, both encoder and decoder paths in MPAM are divided into \( S \) blocks, and each block contains 3 convolution layers with \( 3 \times 3 \) 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 detailed structure of the basic MPAM is shown as the third part of the bottom branch in Fig. 4.

Let’s denote the input and output of the \( s-th \) order scale convolution block in the encoder path as \( X_i^o \) and \( Y_i^o \), respectively, the relation of the input and output of the \( s-th \) scale block can be expressed as:

\[
X_i^o = F_{s,Com}^E (X_i^e, \theta_{s, X}) \quad Y_i^o = F_{s,Com}^D (Y_i^1, \theta_{s, Y})
\]  

(2)

where \( F_{s,Com}^E(\cdot) \) and \( F_{s,Com}^D(\cdot) \) denote the transformation function of 3 convolution-activation layers in the Encoder and decoder with the learned parameters \( \theta_{s, X} \) and \( \theta_{s, Y} \), respectively. The input of the \((s+1) - th \) scale convolution block in the encoder path is a learned down-sampled version from the output \( X_i^o \) of the \( s-th \) scale block via the proposed SCAP, and it is expressed as:

\[
X_i^{s+1} = F_{SCAP}(X_i^o)
\]

(3)

where \( X_i^{s+1} \) has the half size in spatial direction and double channel number of \( X_i^o \). 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 decoder path, which is formulated as:

\[
Y_i^s = F_{Cat}(X_o^s, F_{Up}(Y_o^{s+1}))
\]

(4)

where \( F_{Cat}(\cdot) \) represents the concatenation operation for aggregating the features maps in the corresponding scale of the encoder path and the previous larger scale of the decoder path while \( F_{Up} \) is the simple bilinear up-sample operation. From Eqs. 3 to 8, the feature maps between the adjacent scale blocks of the encoder path are connected using the SCAP module, and the input feature maps in the decoder path are formulated by progressively aggregating the feature in the previous larger scale and the features of the corresponding scale in the encoder. Thus, the final output \( Y_o^1 \) of the first-order scale convolution block in the decoder path would possess the aggregated multi-scale contexts with diverse receptive fields. At last, \( Y_o^1 \) is up-sampled to the resolution of the 0-order scale for augmenting the complementary representation of the backbone subnet, which is further fused with learned feature maps \( X^m \) of the optional middle-phase module as:

\[
\tilde{X}^m = F_{Cat}[X^m, F_{Up}(Y_o^1)]
\]

(5)

where \( \tilde{X}^m \) represents the input to the late-phase module of the backbone subnet.

2) The Lightweight MPAM

As mentioned above, the channel number of the feature maps in the adjacent large scale of both encoder and decoder paths would be doubled, and thus the parameter number of the vanilla convolution layer with the same kernel size (\( 3 \times 3 \)) correspondingly increases to 4 times of the number in the previous scale. With the adopted increasing scales from 0 (only the backbone subnet) to 3 in our MCADN, the parameter numbers are varied as 0.34, 0.72, 2.26, and 8.31MB, respectively. In this section, we aim to exploit a lightweight context hallucinate block (CHB), shown as the fourth part of the bottom branch in Fig. 2, for replacing the vanilla convolution block as the large scales of our proposed MPAM to decrease the model size. Specifically, the proposed CHB is split into several parts, where the first part involves the vanilla convolution layer but with a rigorously controlled channel number while other parts conduct a series of simple linear operations on the learned feature map of the first part to hallucinate more feature maps. Then, all generated features are concatenated to produce the overall feature maps with the same channel number of the vanilla convolution layer.

In detail, given the feature map \( X^{ln} \) with channel number \( L \), we first employ a vanilla convolution layer with kernel size \( 3 \times 3 \) while the reduced channel number: \( \frac{L}{N} \) to learn the intrinsic feature map and formulate it as:

\[
\tilde{X} = F_{Vconv}(X^{ln})
\]

(6)

With the first part of the convolution layer, the parameter number is also reduced to \( \frac{L}{N} \) time due to the decreased channel number. Then, we leverage \( (N-1) \)-groups of depthwise convolution layers with kernel sizes \( k_1, k_2, \ldots, k_{N-1} \) to conduct linear transformation on the intrinsic features for hallucinating more contexts, expressed as:

\[
\tilde{X}_1 = F_{DW}^{k_1}(\tilde{X}), \tilde{X}_2 = F_{DW}^{k_2} (\tilde{X}), \ldots, \tilde{X}_{N-1} = F_{DW}^{k_{N-1}}(\tilde{X})
\]

(7)

Finally, all hallucinated and intrinsic features are concatenated to generate the output of the CHB \( X^{Out} \) with the channel number \( L \), which is formulated as:

\[
X^{Out} = F_{Cat}[\tilde{X}, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{N-1}]
\]

(8)

C. ARTIFACT-ATTENUATING POOLING AND ACTIVATION FUNCTIONS

This part aims to exploit the artifact-attenuating pooling and activation functions for mining effective and important rain-free contexts, and in detail propose the spatial context-aware pooling and activation layers in our network.

1) Spatial Context-Aware Pooling

As introduced in section II the feature maps in the \( s-order \) scale of the MPAM’s encoder path are required to be spatially reduced (usually half-size) for extracting representative context in larger receptive fields. The generic method for
which is formulated as follows:

\[ \hat{X} = \frac{3}{H} \times \text{operation on} \quad \text{kernel size } 3 \times \text{scale}, \]

we further exploit a depth-wise convolution layer with Fig. 2. the SCAP is shown as the fifth part of the bottom branch in maps with our proposed SCAP. The conceptual structure of where \( \hat{X} \) size reduced maps. Specifically, with a feature map \( \hat{X} \in \mathbb{R}^{W \times H \times C} \), 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 the most salient activation such as giving one maximum of 4 activations in a 2 \( \times \) 2 local region. Although the compact feature reflects the active status of a local region, it may be activated by 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 \( \times \) 2 and stride 2, which aggregates the features of the local region with learnable weights into one representative activation, and produce a compact representation \( \hat{X} \in \mathbb{R}^{W/2 \times H/2 \times C} \) as follows:

\[ \hat{X} = f_{2 \times 2}(X) \quad (9) \]

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

\[ \hat{X} = \max(\hat{X}, f_{DW}(\hat{X})) \quad (10) \]

where \( \hat{X} \in \mathbb{R}^{W/2 \times H/2 \times C} \) is the resulted compact feature maps with our proposed SCAP. The conceptual structure of the SCAP is shown as the fifth part of the bottom branch in Fig. 2.

2) Spatial Context-Aware Activation

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 a critical aspect to affect deraining performance. Given the feature map \( X^c \) 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:

\[ X^{ReLU} = \max(X^c, 0) \quad (11) \]

where \( 0 \) denotes a matrix of the same size with \( X^c \) and all elements are 0. To incorporate the surrounding context into consideration, our SCAA firstly adopts a depth-wise convolution on \( X^c \) to produce spatial context aggregated feature map: \( f_{DW}(X^c) \), and then conduct element-wise max operation on \( X^c \) and \( f_{DW}(X^c) \), as shown as the first part of the bottom branch in Fig. 2. The formula of the SCAA is expressed as

\[ X^{SCAA} = \max(X^c, f_{DW}(X^c)) \quad (12) \]

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.

IV. EXPERIMENTAL RESULTS

In this section, we will conduct extensive experiments to evaluate the effectiveness of our proposed multi-scale context aggregation deraining network. We first present the experimental setting including the used datasets, evaluation metrics, and detailed implementation, and then provide the comparisons with the state-of-the-art deraining methods and ablation study.
A. EXPERIMENTAL SETTING

1) Datasets
We have conducted experiments on four deraining datasets: Rain1200 [6], Rain200L [8], Rain200H [8], Rain800 [68] and Rain14000 [5]. Rain1200 dataset contains 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 Rain200L dataset includes images with light rain and is the relatively easy dataset. The training subset contains 1800 image pairs and the test subset has 200 images. Meanwhile, Rain 200H dataset has the same number of training and testing images but being contaminated by more heavy rain with different shapes, directions, and sizes, and this is the most challenging dataset in the deraining community. Rain800 consists 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 streaks to the clean images following the guidelines mentioned in [24] and have fine-grained streaks with noise-like structures. Rain14000 dataset contains 11200 training images and 2800 test images. The images in Rain14000 are generated by artificially adding 14 rain types with different streak directions to each single clean image. Some synthesized rainy images from all five datasets are shown in Fig. 3.

2) Evaluation Metrics
We employ two commonly used evaluation metrics: i.e. peak signal to noise ratio (PSNR) and structure similarity index (SSIM [69]) to assess the performance 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.

3) Training Details
We implemented our deraining CNN model using Keras with TensorFlow backенд. In the training step, we crop 256 × 256 image regions from the training samples and employ Adam [70] to optimize our network. The networks are trained with 500 epochs and the learning rate is set as 2 × 10^{-4}. The simple MSE between the network predictions and the ground-truth clear images is adopted as the loss function for network training.

B. COMPARISONS WITH STATE-OF-THE-ART
We compare our proposed MCADN with the state-of-the-art (SoTA) methods, including deep detail network (DDN) [5], density-aware deraining (DIDMDN) [6], rain convolutional dictionary network (RCDNet) [65], recurrent squeeze-and-excitation context aggregation net (RESCAN) [7], uncertainty guided multi-scale residual learning network (UMRL) [66], progressive deraining network (PreNet) [46], spatial attentive network (SPANet) [67], multi-scale progressive fusion network (MSPFN) [29], and multi-level context gating knowledge transfer network (MCGKT-Net) [9], and multi-stage progressive restoration network (MPRNet) [51].

The deraining models are separately trained with the training pairs in the datasets: Rain200H, Rain200L, Rain800, Rain1200, and Rain14000, and the quantitative metrics are computed with the learned model under the corresponding dataset, respectively. The average quantitative metrics of our MCADN and the compared SoTA methods on all testing images are illustrated in Table 1. It is observed that our proposed MCADN has achieved the highest SSIM and PSNR in all datasets. Compared with the SoTA methods, our approach obtains a great improvement over most methods. Moreover, the visualization examples with our network and different SoTA methods have been shown in Fig. 4. From Fig. 4, we can see that the proposed model can restore clearer results.
TABLE 1. Average PSNR and SSIM comparison on the synthetic datasets Rain1200, Rain200L, Rain200H, Rain800 and Rain14000. Red and blue colors are used to indicate top 1st, 2nd performance.

| Methods | Rain1200 [6] | Rain200L [8] | Rain200H [8] | Rain800 [72] | Rain14000 [5] | # Parameters |
|---------|--------------|--------------|--------------|-------------|-------------|-------------|
| PSNR↑  | 23.64        | 0.727        | 26.71        | 0.834       | 13.79       | 0.367       | 22.18        | 0.663       | 25.16        | 0.782       | -            |
| SSIM↑   |              |              |              |             |             |             |              |              |             |             |             |
| DAIN     | 23.59        | 0.832        | 34.46        | 0.957       | 26.11       | 0.792       | 22.78        | 0.803       | 24.73        | 0.845       | 58.175       |
| JORDER   | 24.32        | 0.862        | 36.95        | 0.979       | 22.05       | 0.727       | 22.24        | 0.776       | 29.03        | 0.888       | 4109.024     |
| DD-MDN   | 29.66        | 0.899        | 35.40        | 0.962       | 26.61       | 0.824       | 22.55        | 0.812       | 29.65        | 0.901       | 135800       |
| RCDNet   | 29.81        | 0.859        | 35.28        | 0.971       | 26.18       | 0.836       | 24.59        | 0.821       | 33.04        | 0.946       | 316635       |
| RESCAN   | 30.54        | 0.879        | 36.09        | 0.970       | 26.75       | 0.835       | 24.99        | 0.830       | 30.51        | 0.882       | 499668       |
| UMRL     | 30.57        | 0.909        | 33.83        | 0.957       | 23.01       | 0.744       | 24.99        | 0.869       | 27.82        | 0.892       | 984356       |
| PreNet   | 31.49        | 0.910        | 37.25        | 0.978       | 28.57       | 0.857       | 24.79        | 0.849       | 30.73        | 0.917       | 168963       |
| SPANes   | 31.94        | 0.902        | 35.79        | 0.965       | 26.27       | 0.865       | 22.41        | 0.838       | 28.46        | 0.880       | 283716       |
| MCPPN    | 32.06        | 0.913        | 31.64        | 0.925       | 27.59       | 0.843       | 27.01        | 0.851       | 32.91        | 0.916       | 15823424     |
| MCGK-T-Net [9] | 32.91 | 0.916 | 37.13        | 0.973       | 28.71       | 0.873       | 28.73        | 0.868       | 33.26        | 0.942       | 14761155     |
| MPNNet  [51] | 32.94 | 0.914 | 35.72        | 0.962       | 29.49       | 0.887       | 29.61        | 0.874       | 33.64        | 0.938       | 3637249      |
| LC-MCDN (This work) | 34.33 | 0.931 | 39.44        | 0.985       | 29.69       | 0.900       | 29.66        | 0.897       | 33.30        | 0.938       | 8306051      |
| LW-MCADN (This work) | 34.76 | 0.937 | 39.54        | 0.986       | 29.92       | 0.904       | 29.70        | 0.876       | 33.80        | 0.949       | 4200147      |

Moreover, to construct a common model with high generalization capability, we further combined different types of rain image datasets to train our LW-MCADN, named LW-MRADN (Common) and then adopt it for recovering the clear images for a real dataset (15 images) [8]. The compared blind perceptual metrics: NIQE, BRISQUE, PIQE with the SoTA models: PReNet [46], MCGK-T-Net [9], MPRNet [51], and MSPFN [29] are given in Table 2, which demonstrates our proposed model is able of achieving the best perceptual results. The compared computational complexity: the required FLOPs in different models are also illustrated in Table 2.

TABLE 2. Comparison of the perceptual evaluation metrics and the computational FLOPs with state-of-the-art methods on a real image dataset [8].

| Method | Input | PReNet [46] | MCGK-T-Net [9] | MPRNet [51] | MSPFN [29] | LW-MCADN |
|--------|-------|-------------|----------------|-------------|-------------|-----------|
| NIQE↑  | 1.758 | 1.856       | 1.848          | 1.817       | 1.794       | 1.816     |
| BRISQUE↑| 0.186 | 0.204       | 0.201          | 0.205       | 0.208       | 0.206     |
| PIQE   | 4.46  | 4.68        | 4.68           | 4.72        | 4.70        | 4.71      |
| FLOPs (10^13) | (512x512) | 2.8 | 4.3 | 5.6 | 24.2 | 3.2 |

C. ABLATION STUDIES

In this section, we evaluate the effectiveness of the different proposed modules including the usage with/without the optional middle-phase module in the backbone residual subnet, the proposed SCAP and SCAA, the MPAM module with different scales, and the lightweight MPAM (LW-MPAM) on Rain 200H dataset. Firstly, we conducted the experiments using the basic MCADN (the backbone subnet without the optional module and with the MPAM of three scales) on both Rain 200L and 200H datasets, and illustrates the compared quantitative values in Table 3 using the proposed SCAP/SCAA or the conventional ReLU activation and the max pooling layer for feature map down-sampling in MPAM module. Table 3 manifests 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. Then, we give the compared results for verifying the effectiveness of different proposed modules. Especially, we denote the model of the backbone residual subnet (the top branch in Fig. 2) without the optional middle-phase module as Baseline (s=0, w/o), with the optional module as Baseline (s=0, w), and the models with the MPAM of different scales as MPAM (s=·, w/o w). Table 4 provides the compared results, and validates that the aggregation with large scale in MPAM is capable of gradually boosting the restoration performance, where the proposed MPAM with scale s = 3 lift the PSNR about 1.71 and SSIM about 0.032. In the meantime, the lightweight version via replacing the vanilla convolution with the proposed context hallucinate block can not only reduce more than half parameters in the original version but also improve the performance. Moreover, we also intuitively employed the separable convolution layer (the combined point-wise and depth-wise convolution layers) instead of the vanilla convolution in MPAM to reduce the model parameters as shown in Table 4.

TABLE 3. Ablation study of the SCAP and SCAA.

| SCA | SCAP | # Parameters (M) |
|-----|-----|----------------|
|     |     | 8.00          | 8.03 | 8.27 | 8.31 |

FIGURE 5. Ablation study of the lightweight modules on the Rain 200H dataset.
parameters, and implemented a simple lightweight version of the MCADN, denoted DWSepConv. Fig. 5 shows the compared results, which manifests that the simple lightweight version with the DWSepConv really reduces the parameters by about 40% but leads to a little performance degradation. While our proposed CHB is capable of performing better on both model size and recovering accuracy.

V. CONCLUSION

This study proposed a novel multi-scale context aggregation network (MCADN), to effectively solve the single image deraining problem. Specifically, we exploited a lightweight subnet with several residual blocks as the backbone architecture to reduce the model size and multi-scale progressive context aggregation module (MPAM) to automatically learn the complementary semantic context for enhancing the modeling capability of the overall deraining network. To reduce the learnable parameters in the MPAM, we explored a context hallucinate block for replacing the vanilla convolution layer, and constructed a lightweight MCADN. Furthermore, we delved into an artifact-attenuating pooling and activation method taking into consideration of the surrounding spatial contexts instead of pixel-wise operation, and proposed the spatial-context-aware pooling (SCAP) and activation (SCAA) for incorporating with our deraining network to boost performance. Extensive experiments on four benchmark datasets demonstrated that our proposed method performed favorably against state-of-the-art deraining approaches on both model size and deraining accuracy.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6479–6488.

[2] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, “Unsupervised vehicle re-identification with progressive adaptation,” in Proc. 29th Int. Joint Conf. Artif. Intell., Jul. 2020, pp. 913–919.

[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. (ICCV), Dec. 2015, Art. no. 108124.

[4] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, “Meta-learning based image decomposition with applications to single image denoising,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2726–2734.

[5] L. Zhu, C.-W. Fu, D. Lischinski, and P.-A. Heng, “Joint bi-layer optimization for single-image rain streak removal,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2526–2534.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NeurIPS, 2012, pp. 1097–1105.

[7] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, “Gather-excite: Exploiting feature context in convolutional neural networks,” in Proc. NeurIPS, 2018, pp. 2–15.

[8] X. Wang, R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Feature rich hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[9] K. Yamamichi and X.-H. Han, “Multi-level context gating knowledge distillation for single image deraining,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 566–575.

[10] K. Yamamichi, X.-H. Han, and X. Wang, “Deep joint rain detection and removal from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1357–1366.

[11] 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.

[12] 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. Multimedia, vol. 16, no. 1, pp. 83–93, Jan. 2014.

[13] 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.

[14] Y. Chang, L. Yan, and S. Zhong, “Transformed low-rank model for line pattern noise removal,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1726–1734.

[15] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, “Clearing the skies: A deep network architecture for single-image rain removal,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7794–7803.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. ECCV, Sep. 2018, pp. 3–19.

[17] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3146–3154.

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 603–612.

[19] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241.

[20] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, “Clearing the skies: A deep network architecture for single-image rain removal,” IEEE Trans. Image Process., vol. 26, no. 6, pp. 2944–2956, Jun. 2017.

[21] X. Fu, J. Huang, Z. Zhang, L. Feng, and N. Xiao, “Gradual network for single image de-raining,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 1795–1804.

[22] Y. Zheng, X. Yu, M. Liu, and S. Zhang, “Residual multiscale based single image deraining,” in Proc. BMVC, 2019, p. 147.

[23] 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. 8346–8355.
KOHEI YAMAMICHI received the B.S. degree from Yamaguchi University, Japan, where he is currently pursuing the master’s degree with the Graduate School of Science and Technology for Innovation. His current research interests include deep learning and single image deraining. He is a member of the IEICE.

XIAN-HUA HAN (Member, IEEE) received the B.E. degree from Chongqing University, Chongqing, China, the M.E. degree from Shan dong University, Jinan, China, and the D.E. degree from the University of Ritsumeikan, Japan, in 2005, From April 2007 to March 2016, she was a Postdoctoral Fellow and an Associate Professor with the College of Information Science and Engineering, Ritsumeikan University, Japan, and from April 2016 to February 2017, was a Senior Researcher with the Artificial Intelligence Researcher Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan. Since March 2017, she has been an Associate Professor with Yamaguchi University, Japan. Her current research interests include image processing and analysis, feature extraction, machine learning, computer vision, and pattern recognition. She is a member of the IEICE.

* * *