IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 August 2022, accepted 18 September 2022, date of publication 26 September 2022, date of current version 6 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210190

RESEARCH ARTICLE

Deep Feature Interactive Aggregation Network for Single Image Deraining

SHAOLI CAO, LIYING LIU, LI ZHAO, YUEWANG XU, JIAWEI XU, AND XIAOQIN ZHANG, (Senior Member, IEEE)
College of Computer Science and Artificial Intelligence, Wenzhou University, Zhejiang 325035, China
Corresponding author: Li Zhao (lizhao@wzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61922064, Grant U2033210, and Grant 62101387.

ABSTRACT Single image deraining aims to remove rain streaks from a degraded input and reconstruct a high-quality image. In recent years, image processing tasks mostly applied a U-shaped architecture to capture rich contextual information. However, it is difficult to achieve long-range pixel dependencies because of the local receptive field of the convolution operation. In this paper, we propose a deep feature interactive aggregation network for single image deraining to enhance long-range dependencies among features and realize the interaction of information. To fully utilize high-level semantic features, we design a long-range dependency feature aggregation module to significantly improve the representational ability of the original U-shaped architecture. It aggregates multi-scale features and calculates the interactive attention of non-overlapping patches among feature maps. In addition, we adopt group normalization to retain the independence of each given image. It interacts with the information among features in an individual image and normalizes the channels of each group to weaken the correlation between batch data processing. Experimental results on widely acknowledged datasets also demonstrate the superiority of our proposed network over previous state-of-the-art methods.

INDEX TERMS Deep network, image deraining, transformer.

I. INTRODUCTION

Rain is a type of severe weather that degrades the quality of images and hampers the application for other vision tasks [1], [2], [3]. Therefore, image deraining has become an important research topic that aims to remove rain streaks from rain images. It can provide better image data for other visual tasks and reduce the interference of low-quality images in model predictions, such as pedestrian trajectory prediction, object tracking, and image classification.

Traditional rain removal methods [4], [5] mainly adopt model-driven algorithms such as Gaussian distribution, sparse coding, and dictionary learning. They utilized the physical properties of rain and prior knowledge of the background in an optimization function and constructed a specific prior model to solve the rain removal problem. Although these models address specific problems, their performance is limited because the prior information is unsuitable for different scenarios. Recently, rain removal tasks have mostly been based on data-driven deep learning methods [6], [7]. Yang *et al.* [6] proposed a region-dependent rain image model to detect rain regions with the aim of solving for heavy rain and rain accumulation. Wang *et al.* [7] adopted a U-shaped network and a residual learning branch to reuse features at different scales. However, owing to the local receptive field of the convolution operation, these methods cannot establish long-range pixel dependencies, resulting in the limitation of the feature interaction. Inspired by the swin transformer [8], we utilize shifted window strategy to model the long-range dependencies. Thus, we aggregate features among different scales to realize the interaction of features and strengthen the detail representation ability of the restored image.

To alleviate the influence of data distribution changes and map data distribution into a specific region, the data-driven

The associate editor coordinating the review of this manuscript and approving it for publication was Eduardo Rosa-Molinar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

deep neural network adopts normalization strategy to accelerate model convergence. Most methods apply batch normalization(BN) [9], which calculates the mean and standard deviation along the batch dimensions of training dataset. However, the use of BN easily leads to the correlation of batch data processing, and it is difficult to realize feature interactions in a single image. Su *et al.* [10] replaced the BN layer to avoid the association between the generated images and batch data processing, then utilized instance normalization(IN) [11] to ensure the independence of each image. Moreover, image restoration tasks often employ small image patches and mini-batches to train the network that easily leads to statistical instability [12]. To solve these problems, we introduce the concept of group normalization(GN) [13]. GN groups along the channel dimension, and calculates the features of each group channel in a single image. Therefore, it avoids the batch axis for feature normalization. In addition, GN is better than IN because it takes advantage of the cross-channel dependencies. Layer normalization(LN) [14] calculates the mean and standard deviation of the channel in each layer, whereas GN learns the channels of each group to obtain different feature distributions. GN significantly improves the interaction ability of features and obtains more characteristics among different data distributions.

In our work, we propose a deep feature interactive aggregation network for single image deraining, realizing the exchange of feature channels and aggregating high-level features to enhance feature dependencies. We design a long-range dependency feature aggregation module consisting of convolutional layers, max pooling layers, and the basic block from swin transformer [8] that provides flexibility for modeling at multiple scales. It has two significant parts. The basic block utilizes the interactive attention mechanism of the non-overlapping shifted window from the aggregation of multiple features. It enhances the pixel dependencies of contextual semantics. A 1×1 convolutional layer is adopted to compress the channels, and the max pooling layer is used to reduce dimensionality. They can eliminate the interference of redundant information and dramatically removes artifacts from the result. We also employ a GN layer instead of a BN layer to weaken the relation between batch data processing. It calculates the feature interaction of a group channel in a single image and significantly increases network performance.

Our contributions are listed as follows.

- We build a long-range dependency feature aggregation module to aggregate the high-level semantic information of the deep network. It achieves interaction of feature information and restores more texture characteristics and color details.
- We utilize the GN layer to perform the normalization of single image feature maps, weakening the correlation between batch data processing and realizing the interaction among features.
- Extensive experiments show that our method performs well in terms of qualitative and quantitative results when compared with state-of-the-art methods.

II. RELATED WORKS

A. SINGLE IMAGE DERAINING TASK

Many classical methods are based on model-driven that restore a rain-free background scene from a given rain image. Liet *et al.* [4] introduced Gaussian mixture model to retain the background information and remove rain streaks. To estimate the rain distribution, Zhu *et al.* [15] analyzed the local gradient statistics of the rain streak direction. Chen *et al.* [16] proposed a generalized low-rank appearance model to capture spatiotemporally correlated rain streaks. Kang *et al.* [5] decomposed the image into low- and high-frequency parts, removing the rain component from the high-frequency part by dictionary learning and sparse coding. Although these methods have achieved good progress in rain removal tasks, they always rely on prior knowledge and lead to the generation of over-smoothed details.

Deep learning brings many advances to image processing methods within the scope of rain removal, as demonstrated in [17], [18], [19]. Fu *et al.* [17] divided the image into the high-frequency detail layer and base layer. They applied ResNet [20] to shorten the mapping range, so as to improve the efficiency of deep network training. To adapt the different types of rain streaks, Zhang *et al.* [18] constructed a classification network to predict the density of rain and employed the category label to guide dissimilar rain removal networks. However, due to the complexity of rain streaks, it is difficult to separate the rain layer from the input image and accurately estimate rain density. Hence, Zheng *et al.* [19] proposed a residual multi-scale pyramid model and used different scales of images as inputs to recover finer details in a coarse-to-fine manner.

Although these methods have achieved fairly achievements under specific conditions, they still have some limitations. They exploited simple ways to transfer feature information at different levels, but it was hard to establish long-range pixel dependencies in the network. Thus, enhancing the interaction capability of features is the most significant research topic in our work.

B. DEEP LEARNING INSPIRED BY TRANSFORMERS

Attention mechanism has made great progress in the field of image restoration [22] and video restoration [23]. Recently, transformer which based on multi-head attention mechanism has broken the predominance of CNN in many computer vision works. Carion *et al.* [24] extracted a compact feature representation from a CNN structure and flattened it into a sequence as the encoder input. It regarded the object detection task as a set prediction problem and output the final sequence of the result directly. Dosovitskiy *et al.* [25] split images into small patches, mapped them into linear embedding sequences, and then fed them into the network to achieve the image classification task. In the work of video super-resolution, VSR-Transformer [26] was proposed and utilized the patch-wise self-attention mechanism to deal with local information. In our work, we aggregate high-level semantic

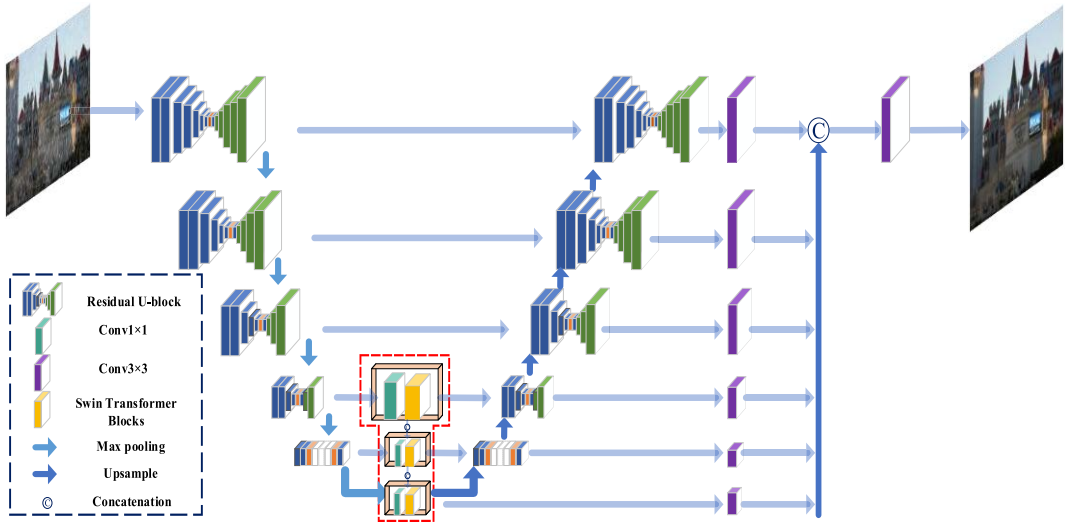

FIGURE 1. The overall architecture of the proposed deep feature interactive aggregation network for single image deraining. The encoder-decoder structure is the Residual U-block of U2-Net [21], and we apply the GN layer in the Residual U-block and replace the BN layer. The long-range dependency feature aggregation module is surrounded by the red dotted line, and its detailed framework is presented in Fig. 2.

features from the convolutional neural network to the transformer, thus enhancing the interaction capability of features and the representation ability of detail information.

III. PROPOSED METHOD

A. OVERALL NETWORK ARCHITECTURE

To make full use of the benefits of high-level features and long-range dependencies, we propose a deep feature interactive aggregation network for single image deraining. The overview framework is shown in Fig. 1, it consists of a U2-Net [21] encoder-decoder structure and a long-range dependency feature aggregation module. The encoder-decoder structure consists of five symmetrical U-shaped subnetworks. Each U-shaped framework is called Residual U-Block, which is stacked by multiple up- and down-sampling operations. This structure can capture the low-level structural feature and high-level contextual information. It is beneficial to restore better image details.

To promote the interaction of image features, we adopt the GN layer as the feature normalization way to replace the BN layer in the Residual U-blocks. In the deep levels of our network, the long-range dependency feature aggregation module is embedded to capture finer detail information. It consists of three steps. First, the input feature utilizes a 1×1 convolutional layer to compress the number of channels from 512 to 96. Then, the different scales of features are aggregated by a concatenation operation, and the feature is puted into a pair of swin transformer blocks after stretching into a three-dimensional vector. Finally, after recovering the feature dimension, max pooling removes redundant information by filtering the feature, and the last convolutional layer is used to restore the number of channels from 96 to 512.

B. LONG-RANGE DEPENDENCY FEATURE AGGREGATION MODULE

The main goal of this module is to establish long-range dependencies among features by aggregating high-level rich semantic information. The feature aggregation of different levels can focus on the global range semantics well and maintain finer structural detail as sharp as the ground-truth(GT) image. Moreover, the operations of compressing the number of channels and reducing the spatial size of features help alleviate the computational load, eliminate the interference of redundant information and effectively relieve the generation of artifacts.

The module structure is presented in Fig. 2, it has three steps. Let f_i denotes the output of the *i-th* Residual U-block. First, a 1×1 convolutional layer is applied to compress the channels of f_4 in which the number of channels is changed from 512 to 96. Subsequently, flattening the feature map from $[B, C, H, W]$ to $[B, H \times W, C]$,

$$\tilde{f_4} = R(\phi(f_4)), \tag{1}$$

where $\phi(\cdot)$ and $R(\cdot)$ express the convolutional layer and flattened vector operation, respectively. The swin transformer blocks are applied to extract finer detail features, where the number of blocks is 2, the window size is set as 7, and the number of self-attention heads is equal to 3. After that, a pooling operation is employed to process the feature map of swin transformer blocks,

$$F_{out_4} = \mathrm{MAP}(R(\beta(\tilde{f_4}))), \tag{2}$$

where $\mathrm{MAP}(\cdot)$ denotes the max pooling operation and $\mathfrak{B}(\cdot)$ refers to a pair of swin transformer blocks. Here, $R(\cdot)$ is used to adjust the size of the feature map from $[B, H \times W, C]$ to $[B, C, H, W]$. Then the output feature F_{out_4} is sent to

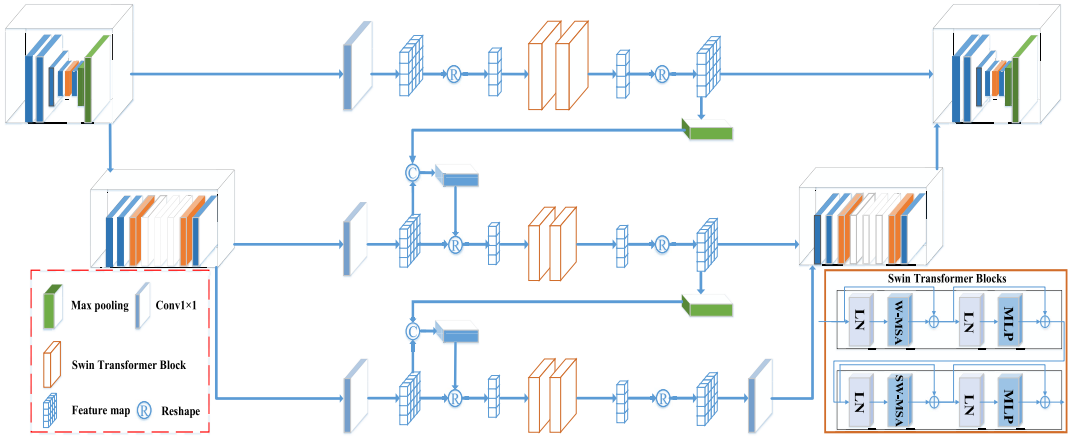

FIGURE 2. Long-range dependency feature aggregation module.

the second step. Through the same operation, we obtain the second output F_{out_5}:

$$F_{out_5} = \text{MAP}(R(\beta(R(\phi(\phi(f_5)©F_{out_4}))))), \tag{3}$$

where © refers to the concatenation operation. The output feature F_{out_4} is aggregated to further extract effective information from the image. The last step repeats operation (3) to obtain the F_{out_6} and restores the number of channels from 96 to 512. Through a series of operations, the module transfers fusion features to the next Residual U-block of the network.

C. GROUP NORMALIZATION FOR DERAINING TASK

We employ the concept of GN in the Residual U-blocks, which avoids the dependence of features among different images and effectively reduces the inaccurate estimation of statistical data. GN divides the feature vectors into multiple groups along the (H, W) axis and a group of C/G channels, where C and G represent the number of channels and groups, respectively. The mean μ and standard deviation σ are calculated from a set of pixels named S_i, and the formulas of μ and σ are listed as follows:

$$\mu_i = \frac{1}{m}\sum_{k\in S_i} xk, \quad \sigma_i = \sqrt{\frac{1}{m}\sum_{k\in S_i}(x_k - \mu_i)^2 + \varepsilon}, \tag{4}$$

where x and m symbolize the layer pixel feature calculation and size of S_i, respectively. ε and $i = (i_N, i_C, i_H, i_W)$ are a small constant and 4D vector index. The difference in normalizations depends on how S_i is determined, where GN is defined as:

$$S_i = \left\{k \mid k_N = i_N, \left\lfloor \frac{k_C}{C/G} \right\rfloor = \left\lfloor \frac{i_C}{C/G} \right\rfloor\right\}, \tag{5}$$

where $\lfloor\cdot\rfloor$ indicates floor operation. In this study, G is set as 4. By replacing the normalization layer, the proposed network effectively avoids the impact of batch data processing estimation and realizes good results of the test dataset.

D. LOSS FUNCTION

We use the combination of the robust Charbonnier loss [27] and edge loss to train the network. They achieve faster convergence in the training and obtain cleaner image details. The Charbonnier loss function can be expressed as:

$$L_{Char} = \sqrt{(I - I^*)^2 + \epsilon^2}, \tag{6}$$

where I and I^* represent the input rain image and corresponding GT image, respectively. ϵ indicates the penalty coefficient, which is set as 10^{-3}. In addition, the edge loss function is expressed as:

$$L_{edge} = \sqrt{(\nabla(I) - \nabla(I^*))^2 + \epsilon^2}, \tag{7}$$

where ∇ expresses the Laplacian operator. The total loss function in this work is formulated as follows:

$$L_{all} = \alpha L_{Char} + \beta L_{edge}. \tag{8}$$

To balance the loss items, α and β set to 1 and 0.05 [28], respectively.

IV. EXPERIMENTS AND ANALYSIS

A. DATASETS

Our proposed network employs extensive synthetic public datasets [4], [6], [17], [29] during training; the number of samples is summarized in Table 1. There are 13,712 pairs of clean/rainy images in the training dataset. Rain14000 [17] contains 11,200 image pairs, which are artificially synthesized using 800 clean images. Each clean image is matched with 14 rainy images with different rain pattern orientations and densities. To verify the effectiveness of the proposed network, we conduct experimental research on five widely used datasets [6], [17], [18], [29]. Specifically, Test1200 [18] is synthesized by Photoshop and has three types of rain density images: light rain, midden rain, and heavy rain. To quantitatively evaluate the methods, we employ commonly used indicators, such as peak signal-to-noise ratio(PSNR) and structural similarity(SSIM).

TABLE 1. Data description. A total of 13,712 clean/rainy image pairs are used for training. Five public datasets are listed and renamed in the last row.

Datesets	Rain12 [4]	Rain100H [6]	Rain100L [6]	Rain800 [29]	Rain1200 [18]	Rain1800 [6]	Rain14000 [17]
Training Samples	12	0	0	700	0	1800	11200
Test Samples	0	100	100	100	1200	0	2800
Test dataset Rename	-	Rain100H	Rain100L	Test100	Test1200	-	Test2800

TABLE 2. Ablation results of long-range dependency feature aggregation module in terms of PSNR, SSIM. The best results are marked in bold.

Datasets		Test1200 [18]		Test2800 [17]		Ave	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
basic model (w/o block)		30.90	0.900	31.36	0.915	31.13	0.908
block numbers	2(Ours)	**32.77**	**0.918**	**32.98**	**0.931**	**32.88**	**0.925**
	4	32.61	0.913	32.76	0.929	32.69	0.921
	6	32.63	0.913	32.74	0.929	32.69	0.921

TABLE 3. Ablation results of normalization layers in terms of PSNR, SSIM. The best results are marked in bold.

Datasets	Test1200 [18]		Test2800 [17]		Ave	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BN	30.17	0.891	29.95	0.892	30.06	0.892
IN	25.42	0.900	24.84	0.906	25.13	0.903
LN	32.27	0.916	32.82	0.930	32.55	0.923
GN(Ours)	**32.96**	**0.921**	**32.97**	**0.931**	**32.97**	**0.926**

B. IMPLEMENTATION DETAILS

The proposed network is trained on NVIDIA Titan-X GPU, and the quantitative results are test on a PC with an Intel Core i5-10500 CPU, 16GB RAM. The model employs Adam optimizer [30], where β_1 and β_2 default to 0.9 and 0.999, respectively. The initial learning rate is set as 4×10^{-4}, and it will decay to 1×10^{-6}. In the training process, the number of iterations is set as 500 epochs, the batch size is fixed to 32, and the group of GN is equal to 4. The input image pair is cropped to a size of 256×256 randomly and then used for data enhancement, including flipping and rotating in different dimensions.

C. ABLATION EXPERIMENT

1) VALIDATION ON LONG-RANGE DEPENDENCY FEATURE AGGREGATION MODULE

The long-range dependency feature aggregation module has the significant value in the proposed network, so we analyze its impact in this section. We define a basic model that removes any swin transformer block. Based on the basic model, we investigate the optimal number of swin transformer blocks. All the results are presented in Table 2, which

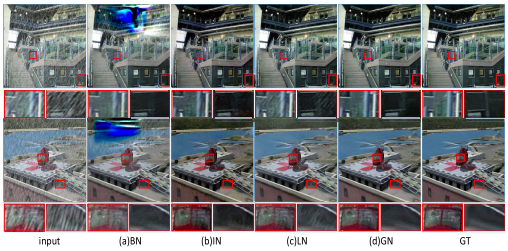

FIGURE 3. Ablation experiments of different normalization layers, where (a)-(d) denote the results of the BN, IN, LN, and GN layers in the network. The network using GN layer achieves the best visualization results than other normalization layers.

contrasts the metrics of PSNR and SSIM on the Test1200 and Test2800 datasets.

The findings are presented as follows. As a basic model, it achieves a reasonable performance. Aggregating high-level features has better results compared to the basic model, which verifies the effectiveness of our proposed module. However, with an increase in the number of blocks, the effect of the model does not exhibit an enhanced trend. It proves that simply stacking blocks is an ineffective way and easily leads to the complexity in the proposed model. When the number of blocks is set as 2, it can improve the performance by 0.19dB and 0.004 compared with the block numbers of 4 and 6.

2) VALIDATION ON GROUP NORMALIZATION LAYER

The method of normalization has a considerable impact on the proposed network; thus, we analyze its effect on the results. We compare four normalization ways: LN [14], IN [11], BN [9], and GN [13]. The qualitative and quantitative results are shown in Fig. 3 and Table 3, respectively.

Several defects are discovered among other normalization ways. In Fig. 3(a), the restored image has the color spot and retains some rain streaks when the BN is used for normalization. The color spot appears randomly on the test datasets and causes degradation of the image visual quality. Using IN layer leads to the problem of color distortion. The overall tone of the restored image is darker in Fig. 3(b). When employing the LN layer, it is difficult to solve the problem of image artifacts, and the final image still contains rain streaks in some areas. To avoid the above drawbacks, we adopt GN layer and it provides comfortable visual effects. Obviously, it has a higher accuracy than other normalization layers, which illustrates that our strategy is the most suitable.

D. COMPARISONS WITH STATE-OF-THE-ART METHODS

1) QUANTITATIVE RESULTS

To verify the superiority of the proposed network, we compare it to other state-of-the-art methods. The evaluation

TABLE 4. Comparison results with several SOTA methods on the different benchmarks. The best and the second best results are marked in bold and underline, respectively.

Datasets	JORDER [6]		RESCAN [31]		SEMI [32]		PReNet [33]		OUCD [34]		ECNet [35]		Ours	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Rain100H [6]	27.75	0.840	26.69	0.812	18.08	0.577	27.49	0.863	24.38	0.733	27.91	0.866	**29.48**	**0.876**
Rain100L [6]	31.27	0.923	30.89	0.919	22.25	0.842	32.91	0.953	29.84	0.902	33.42	0.955	**35.20**	**0.956**
Test100 [29]	24.72	0.850	25.46	0.851	20.72	0.687	24.94	0.853	23.58	0.805	27.55	0.885	**28.90**	**0.887**
Test1200 [18]	31.27	0.903	30.75	0.888	23.91	0.716	31.41	0.911	26.09	0.827	30.05	0.901	**32.92**	**0.920**
Test2800 [17]	31.60	0.919	31.49	0.909	24.38	0.736	31.87	0.927	28.72	0.891	32.42	**0.932**	**33.12**	**0.932**
Ave	29.32	0.887	29.06	0.876	21.87	0.712	29.72	0.901	26.52	0.832	30.27	0.908	**31.92**	**0.915**

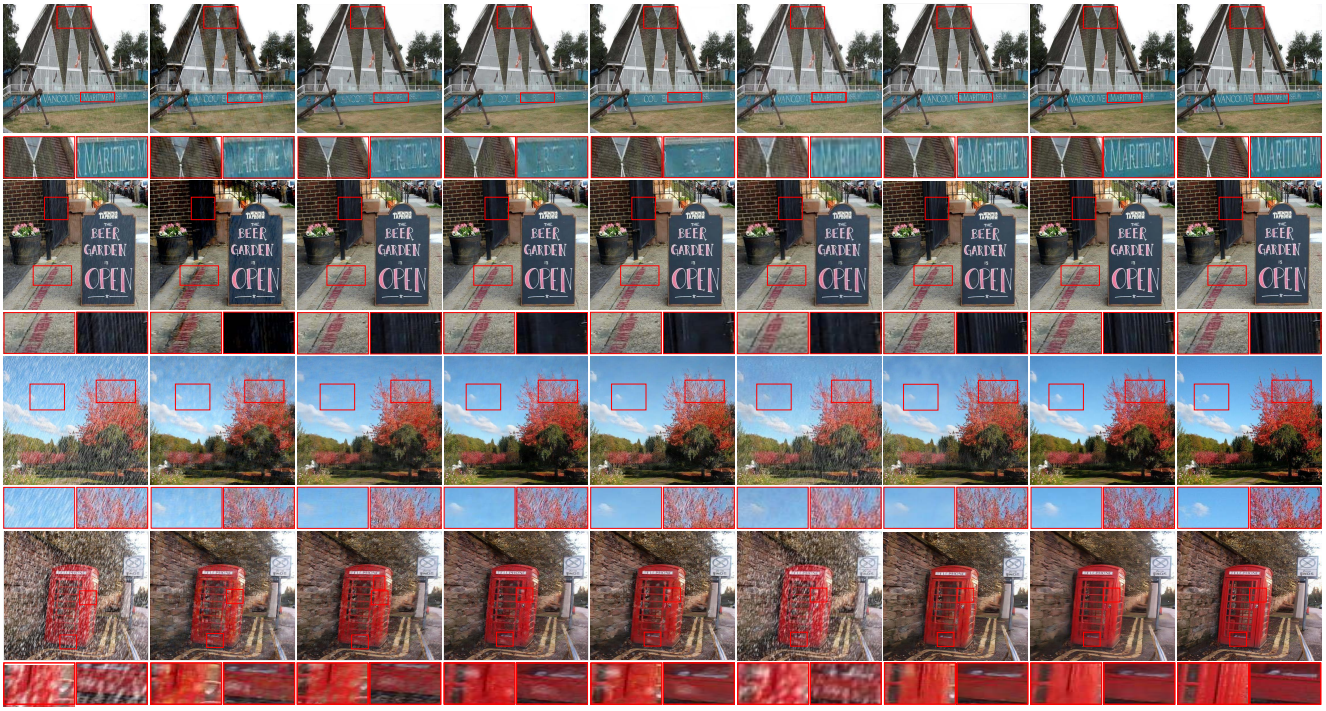

FIGURE 4. Qualitative visual results on synthetic datasets. Compared with other state-of-the-art methods, our proposed model generates the best visualization results by effectively removing the rain streaks and preserving the finer structure derails.

results are listed in Table 4. The compared methods include JORDER [6], RESCAN [31], SEMI [32], PReNet [33], OUCD [34], and ECNet [35]. The published codes are used for the same training and test datasets to obtain fair evaluation results. Specifically, SEMI employs a semi-supervised strategy; ECNet utilizes a rain-to-rain autoencoder to reconstruct an ideal rain embedding, and the evaluation results prove its effectiveness. The results prove that the proposed method has significant progress compared to SOTA methods in terms of both the average PSNR and SSIM metrics. Compared with the second-ranking method ECNet, our method increases by an average of 1.65dB and 0.007.

2) QUALITATIVE RESULTS

We select several representative images from light, middle, and heavy rain, as shown in Fig. 4. Most methods retain large region artifacts, which quite degrade the quality of the restored image. It can be observed that this situation appears in the sky, clouds and roof in SEMI, RESCAN and JORDER. Additionally, because the color of the rain streaks is similar to the background, some methods easily lead to excessive deraining. They remove the details with similar colors together, such as the white font in the first row in Fig. 4. For some dense objects, it is difficult to recover the fine details and remove the rain streaks simultaneously, such

as the black fence and telephone booth in the second and fourth rows in ECNet, SEMI, PReNet and JORDER. OUCD pays more attention to local features and combines global information in the network, but its capability to remove rain streaks is poor in heavy rain. Compared with these methods, our proposed network can avoid the above problems, and the restored images are highly similar to the GT images.

V. CONCLUSION

In this paper, we introduce a new network framework called deep feature interactive aggregation network to address the limitations of the local receptive field and build up feature interactions. A long-range dependency feature aggregation module is designed to improve representation ability and restore better texture details. To realize the interaction of multiple channel information, we adopt GN to normalize the feature maps. The experimental results demonstrate the superiority of the proposed network by comparing it with several state-of-the-art methods.

In future work, we plan to explore a general processing model for image restoration tasks, including but not limited to image deblurring, denoising, and dehazing. This model simplifies the image restoration problem and has become a research topic in recent years.

REFERENCES

[1] K. Duan, L. Xie, H. Qi, S. Bai, Q. Huang, and Q. Tian, "Corner proposal network for anchor-free, two-stage object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 399–416.
[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[3] X. Zhang, W. Hu, S. Chen, and S. Maybank, "Graph-embedding-based learning for robust object tracking," *IEEE Trans. Ind. Electron.*, vol. 61, no. 2, pp. 1072–1084, Feb. 2014.
[4] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2736–2744.
[5] 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. 2011.
[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1357–1366.
[7] G. Wang, C. Sun, and A. Sowmya, "ERL-Net: Entangled representation learning for single image de-raining," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5644–5652.
[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[10] Z. Su, H. Zhang, Z. Liu, and X. Zhu, "Image deblurring algorithm by using conditional generation adversarial network," in *Proc. 40th Chin. Control Conf. (CCC)*, Jul. 2021, pp. 8128–8133.
[11] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.
[12] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," 2018, *arXiv:1808.08718*.
[13] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[14] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[15] 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.
[16] 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. 2013, pp. 1968–1975.
[17] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3855–3863.
[18] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 695–704.
[19] Y. Zheng, X. Yu, M. Liu, and S. Zhang, "Residual multiscale based single image deraining," in *Proc. BMVC*, 2019, p. 147.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[21] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.
[22] X. Zhang, J. Wang, T. Wang, and R. Jiang, "Hierarchical feature fusion with mixed convolution attention for single image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 510–522, Feb. 2022.
[23] X. Zhang, T. Wang, R. Jiang, L. Zhao, and Y. Xu, "Multi-attention convolutional neural network for video deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1986–1997, Apr. 2022.
[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.
[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[26] J. Cao, Y. Li, K. Zhang, and L. Van Gool, "Video super-resolution transformer," 2021, *arXiv:2106.06847*.
[27] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, Nov. 1994, pp. 168–172.
[28] 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.
[29] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3943–3956, Nov. 2020.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[31] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 254–269.
[32] W. Wei, D. Meng, Q. Zhao, Z. Xu, and Y. Wu, "Semi-supervised transfer learning for image rain removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3877–3886.
[33] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3937–3946.
[34] R. Yasarla, J. M. J. Valanarasu, and V. M. Patel, "Exploring overcomplete representations for single image deraining using CNNs," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 2, pp. 229–239, Feb. 2021.
[35] Y. Li, Y. Monno, and M. Okutomi, "Single image deraining network with rain embedding consistency and layered LSTM," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 4060–4069.

SHAOLI CAO received the B.E. degree in digital media technology from Guangdong Polytechnic Normal University, China, in 2019. She is currently pursuing the Graduate degree in computer software and theory with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China. Her research interests include computer vision and machine learning, such as image quality restoration and salient object detection.

LIYING LIU received the B.Sc. degree from the Department of Information and Computing Sciences, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2020. She is currently pursuing the Graduate degree in computer software and theory with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China. Her research interests include image processing and deep learning.

LI ZHAO received the B.Sc. degree in automation in 2005 and the M.Eng. degree in control theory and control engineering from Central South University, China in 2008. She is currently an Assistant Researcher with Wenzhou University. Her research interests include pattern recognition, computer vision, and machine learning.

YUEWANG XU received the B.Sc. degree in computer science and technology from the Shandong University of Technology, China, in 2020. He is currently pursuing the Graduate degree with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. His research interests include several topics in computer vision and machine learning, such as salient object detection, image/video super-resolution, and recommendation algorithm.

JIAWEI XU received the Ph.D. degree with the topic of human factors in driving from the University of Lincoln, U.K., in 2015. He has been an Adjunct Consultant at Huawei Technologies specialized in human factors in driving and an Associate Professor at Wenzhou University, since 2019. He was a Postdoctoral Researcher and a Lecturer at Newcastle University, U.K., from 2015 to 2019. His research interests include human factors in aviation and human factors in intelligent/automated driving.

XIAOQIN ZHANG (Senior Member, IEEE) received the B.Sc. degree in electronic information science and technology from Central South University, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China, in 2010. He is currently a Professor with Wenzhou University, China. He has authored or coauthored over 100 papers in international and national journals and international conferences. His research interests include pattern recognition, computer vision, and machine learning.