IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 July 2022, accepted 21 August 2022, date of publication 29 August 2022, date of current version 7 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202888

RESEARCH ARTICLE

Single Image Raindrop Removal Using a Non-Local Operator and Feature Maps in the Frequency Domain

SHINYA EZUMI AND MASAAKI IKEHARA, (Senior Member, IEEE)
Department of Electronics and Electrical Engineering, Keio University, Yokohama 223-8522, Japan
Corresponding author: Shinya Ezumi (ezumi@tkhm.elec.keio.ac.jp)

ABSTRACT Taking a photo on a rainy day may result in a photo with raindrops. Images containing raindrops have a significant impact on the visual impression and accuracy when applied to image recognition systems. Thus, an automatic high-quality raindrop removal method is desired for outdoor image processing systems as well as for acquiring good-looking images. Several existing methods have been proposed to tackle this problem, but they often fail to keep global consistency and generate unnatural patterns. In this paper, we tackle this problem by introducing a non-local operator. The non-local operator combines features in distant locations with matrix multiplication and enables consistency in distant locations. In addition, high-frequency components such as edges are more affected in images with raindrops. Inspired by the nature that high-frequency components can be separated from other components in the frequency domain, we also propose to process feature maps in the frequency domain, which are obtained by the fast Fourier transform operation and processed by several convolution layers. Experimental results show that our method effectively removes raindrops and achieves state-of-the-art performance.

INDEX TERMS Deep learning, fast Fourier transforms, image processing, image restoration.

I. INTRODUCTION

Nowadays, outdoor image processing systems are widely used in many devices such as driving support systems and surveillance cameras. On rainy days, cameras used for these systems often get images with raindrops. Similar situations will occur when cameras are indoors. Water drops are adhered to glass due to condensation or water splash. Raindrops severely hinder the visibility and quality of images, so they can negatively impact visual recognition systems. Therefore, there is a growing demand for automatic raindrop removal systems that can effectively recover areas occluded by raindrops to produce more visually pleasing images.

Recently, several deep learning methods [1], [2], [3] have been proposed for raindrop removal, and these methods use convolution layers as the base structure. The convolution layer has a limited receptive field, which can efficiently and effectively capture local patterns. However, the convolution layer is not good at capturing non-local patterns in

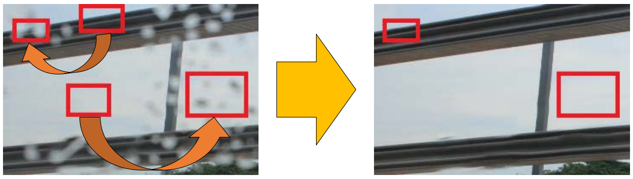


FIGURE 1. Example of raindrop removal using features in distant locations.

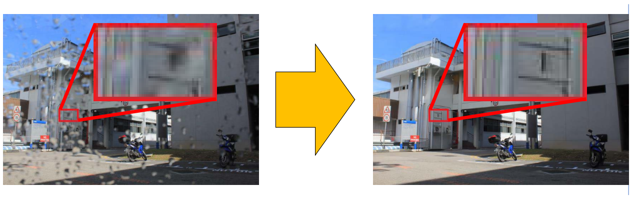


FIGURE 2. Example of inpainting details occluded by raindrops.

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

images. Stacking several layers makes it possible to enlarge the receptive field, but it is still ineffective in capturing long-range dependencies. Therefore, these previous methods cannot capture the long-range dependencies sufficiently and output unnatural artifacts in images.

To solve the above problem, we incorporate a non-local operation in our model which helps capture long-range dependencies. Fig. 1 shows an ideal example of raindrop removal using long-range dependencies. Features in distant locations can be used for effective and efficient raidrop removal. The non-local operator incorporates features in distant places using matrix multiplication of the input feature map and its transpose. In our proposed model, we use the structure of Global Context Network [4] (GCNet) as the non-local operator.

Furthermore, an image can be decomposed into several components such as backgrounds, foregrounds, and edges. As for images with raindrops, high-frequency components such as edges are usually affected more. Therefore, it is effective for raindrop removal methods to process feature maps in multiple resolutions, and thus many methods employ feature maps in multiple resolutions. In these methods, several smaller feature maps are first generated from prior feature maps and then each is processed by an independent unit.

Each component in an image is better separated in the frequency domain than in the spatial domain. For example, sharp edges and background are mixed in the spatial domain, but these components can be completely separated into high and low frequencies. Thus, we propose to process feature maps in the frequency domain to restore high-frequency details more realistically. Fig. 2 shows an ideal example of inpainting details occluded by raindrops. Fine details are often occluded in images with raindrops and need to be repaired precisely for effective raidrop removal.

This study uses Residual FFT Convolution (Res FFT-Conv) block from Deep Residual Fourier Transform [5] (DeepRFT) as the base architecture of a processing unit to utilize feature maps in the frequency domain. Res FFT-Conv block uses the fast Fourier transform (FFT) to generate feature maps in the frequency domain and processes them with several convolution layers.

Additionally, we use two additional loss functions specialized for restoring high-frequency components to ensure more accurate raindrop removal. One uses images in the frequency domain, and the other uses images processed by the Laplacian filter, which can extract edges.

Our contributions are

1) We introduce a non-local operator to solve the problem that previous methods have in keeping consistencies of pixels in distant locations.
2) We introduce feature maps in the frequency domain and loss functions specialized for high-frequency components to remove raindrops and recover the occluded areas more realistically.

Experimental results show that the proposed method is effective for more realistic raindrop removal and achieves competitive performance compared to state-of-the-art methods for raindrop removal.

II. RELATED WORKS

A. IMAGE PROCESSING METHODS FOR BAD WEATHER

There are many methods that tackle images occluded by bad weather, such as haze/fog removal [6], [7], [8], [9] and snow removal [10], [11], [12]. Many deraining methods [13], [14], [15] have also been proposed and showed high performance in rain streak removal. However, these methods cannot effectively restore images with raindrops. This failure occurs due to the shape and size difference between raindrops and other obstacles.

Both video-based (e.g. [16], [17]) and single image-based methods have been proposed for raindrop removal. As for video-based raindrop removal, Roser *et al.* [16] proposes to remove raindrops by detecting raindrops using a photometric raindrop model before repairing the occluded areas using neighboring image frames. You *et al.* [17] proposes to detect raindrops based on the difference of both the motion speed and the intensity between the pixels with and without raindrops. You *et al.* [17] also improved the removal procedure by using either images in neighboring pixels or other frames depending on how severely the area is occluded by raindrops.

Though these methods are successful to some extent in removing raindrops in video images, they require multiple frames. To acquire visually pleasing images only from a single image with raindrops, we adopt a single image-based method. Additionally, strategies for single image-based methods can also be applied to individual video frames.

For raindrop removal from a single image, Eigen *et al.* [18] is the first method that tackled raindrop removal using a couple of CNN layers. Though this method works well on images with small and sparse raindrops, it is ineffective for images with large and dense raindrops.

Qian *et al.* [1] proposes a GAN [19] based method which is supported by an attention map. The attention map is generated by residual blocks [20] and LSTM units [21], and plays the role as a binary mask that shows how much each position is influenced by raindrops. The output clear image is then generated by several convolution layers from the input image and the generated attention map.

Quan *et al.* [2] proposes a shape-driven attention module to utilize characteristics of raindrops' shape such as roundness and closedness to better restore a clean image. An edge map extracted from the input image is also used to support the shape-driven attention. The edge map is extracted based on the difference magnitude.

He *et al.* [3] tackles the problem of restoring images containing both mists and raindrops. This method uses FFA-Net [7], a network proposed for haze removal, as the base structure. To effectively remove raindrops, this method proposes the interpolation-based pyramid attention (IPA) block and adds several IPA blocks to the base structure. The IPA block processes multi-resolution feature maps to capture information and process feature maps more effectively.

All in One [22] is a method that can deal with multiple types of bad weather containing fog, raindrop and snow. All in One [22] adopts adversarial learning and its generator consists of multiple encoders specialized for each bad weather and a decoder generalized for multiple terms. The architecture of the encoder combines various kinds of operators corresponding to several fundamental operations for feature searching.

B. MULTI-RESOLUTION PROCESSING

Many methods for image processing including [1], [2], [3] use multi-scale feature maps and this architecture has shown effectiveness for various tasks such as image classification [32], image super-resolution [33], and motion prediction [34].

He *et al.* [3] introduces multi-scale processing to each processing block and generates feature maps in each resolution by image interpolation. Qian *et al.* [1] and Quan *et al.* [2] introduce multi-scale processing to the whole model. The two methods use U-Net [23] as the base structure of their models.

U-Net [23] is the first method that uses convolution layers for image-to-image processing and has demonstrated better performance in image segmentation tasks than previous methods. Following [24] and [25], U-Net adopts multi-scale processing and feature maps at each resolution which are generated by max pooling. There are two processing groups at each resolution, where one is the encoder block and the other is decoder block. U-Net proposes to improve the multi-scale processing procedure by introducing skip connection: each decoder can utilize not only the up-scaled feature maps of those in the lower resolution but also the feature maps in the current resolution. This procedure allows the network to preserve positional homogeneity which is reduced by the convolution layers.

Multi-input multi-output U-Net [26] (MIMO-UNet) has been proposed to improve some of the problems in U-Net [23]. One of the improvements is the asymmetric feature fusion (AFF) module, which generates the flexible flow of feature information in the network architecture. Another is the multi-output single decoder, which is used for multi-scale losses. We utilize MIMO-UNet [26] as the base structure of our proposed network so that the processing units at each scale can deal with a specific component.

III. PROPOSED METHOD

We design our proposed network with the following concepts.

1) We utilize MIMO-UNet [26] as the base structure so that each structure can deal with a specific component and features at each scale are flexibly processed.
2) We introduce a module which processes feature maps in the frequency domain to more realistically restore high-frequency components such as edges.
3) We add the non-local operator to help capture features in distant locations.
4) We introduce loss functions which are specialized for recovering high-frequency components to ensure precise restoration of high-frequency components.

Fig. 3 shows the overall structure of our proposed network. The same as MIMO-UNet [26], there are two processing groups at $k \in \{1, 2, 3\}$-th resolution, where one is the encoder block (EB_k) and the other is decoder block (DB_k). In Fig. 3, $input_k$ denote the input images whose size are $(H \times W)$, $(H/2 \times W/2)$, and $(H/4 \times W/4)$ for the downsized images, respectively, and $output_k$ denote the output images of the corresponding decoder block, respectively. Refer to [26] for the detailed structure of AFF and Shallow Convolutional Module (SCM) block. In the following, we will describe the details of each component.

A. PROCESSING FEATURE MAPS IN THE FREQUENCY DOMAIN

To restore high-frequency details more realistically, we propose to utilize feature maps in the frequency domain so that each component in the feature map can be processed more independently. Concretely, we introduce Res FFT-Conv block as a basic module in MIMO-UNet. Res FFT-Conv block, proposed in [5], uses the FFT to convert feature maps into the frequency domain and processes them with 1 × 1 convolution layers. The lower left of Fig. 3 shows the architecture of a Res FFT-Conv block. $\mathbf{Z}$ is the input feature maps and $\mathbf{Y}^{\mathbf{res}}$ and $\mathbf{Y}^{\mathbf{fft}}$ are the feature maps from the residual block [20] and the frequency domain, respectively.

In the flow of $\mathbf{Y}^{\mathbf{fft}}$, feature maps in the frequency domain $\mathcal{F}(\mathbf{Z}) \in \mathbb{C}^{H \times W/2 \times C}$ are first generated by the 2D Real FFT from $\mathbf{Z} \in \mathbb{R}^{H \times W \times C}$. H, W, and C denotes height, width, and the number of channels of the input feature maps, respectively. After converting $\mathcal{F}(\mathbf{Z})$ into $\mathbb{R}^{H \times W/2 \times 2C}$ by concatenating its real part and imaginary part, processed feature maps in the frequency domain ($f \in \mathbb{R}^{H \times W/2 \times 2C}$) are generated by processing $\mathcal{F}(\mathbf{Z})$ using two 1×1 convolution layers and a Rectified Linear Unit (ReLU) activation. Finally, the output feature maps $\mathbf{Y}^{\mathbf{fft}}$ are generated by the 2D Real IFFT after converting f into $\mathbb{C}^{H \times W/2 \times C}$ by splitting them in the channel direction by two.

To speed up the convergence, we replace convolution layers in our proposed network by Depthwise Over-parameterized Convolution (DO-Conv [27]) layers except for 1 × 1 convolution layers.

B. NON-LOCAL OPERATION

NLNet [28] introduces the non-local operator inspired by the non-local means filter [29]. The non-local operator calculates each output by a weighted sum of all points in the feature map. The output of a non-local operator at point i in a feature map is expressed as:

$$\mathbf{y}_i = \sum_{\forall j} \frac{1}{C(\mathbf{x})} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j) \tag{1}$$

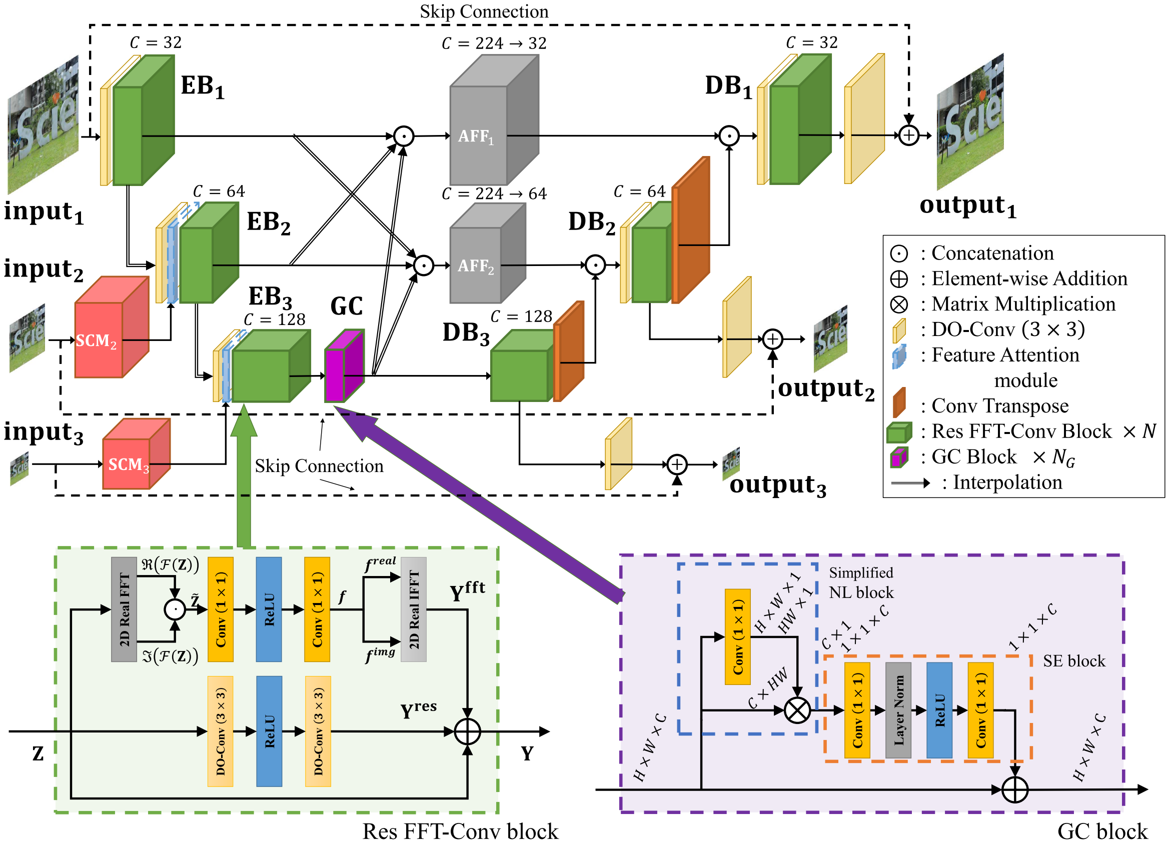


FIGURE 3. Overall architecture of our proposed network.

where j denotes a point in the feature map, $f(\mathbf{x}_i, \mathbf{x}_j)$ denotes the similarity of features at point i and j, and g denotes a function which represents features at j. C is used for normalization.

To solve the problem that existing methods have of not being able to capture long-range dependencies and generating unnatural artifacts, we propose to add a module which can capture long-range dependencies and utilize them for processing. Concretely, we insert Global Context (GC) blocks as the non-local operator into immediately following the encoder block at the lowest resolution. We select this position to effectively capture relationships between features at distant locations while suppressing the increase in computational complexity. The lower right of Fig. 3 shows the architecture of a Global Context (GC) block.

In the previous non-local operation including NLNet [28], $f(\mathbf{x}_i, \mathbf{x}_j)$ is calculated by the matrix product of the input feature maps and the transpose of itself for each channel. The GC block, proposed in [4], calculates $f(\mathbf{x}_i, \mathbf{x}_j)$ only by a convolution layer with channel reduction. The calculation of $g(\mathbf{x}_i)$ is also simplified in the GC block: the same as the input feature maps. The simplification is based on the idea of sharing the same $f(\mathbf{x}_i, \mathbf{x}_j)$ at all locations and suppresses the increase in computational complexity.

In addition, inspired by Squeeze-Excitation (SE) Net [30], GC block adopts channel-wise processing to enhance the effectiveness of the non-local operator. Since the GC block has a simple structure, the increase in the computational cost of our proposed model is suppressed.

C. LOSS FUNCTIONS

We design the loss function used in the training phase by adding loss functions specialized for recovering high-frequency components to restore high-frequency details more realistically. Also, we set all losses as multi-scale loss so that processing blocks in the smaller scale are trained effectively.

Concretely, we define the total loss $\mathcal{L}$ as the weighted sum of three losses: Multi-Scale Charbonnier (MSC) loss, Multi-Scale Edge (MSED) loss, and Multi-Scale Frequency Reconstruction (MSFR) loss.

MSC loss is defined as

$$\mathcal{L}_{msc} = \sum_{k=1}^{3} \sqrt{\|\hat{\mathbf{S}}_k - \mathbf{S}_k\|^2 + \epsilon^2} \tag{2}$$

$\mathbf{S}_k (k \in \{1, 2, 3\})$ denote the ground truth image whose size is $(H \times W)$, $(H/2 \times W/2)$, and $(H/4 \times W/4)$ for the downsized images, respectively. $\hat{\mathbf{S}}_k$ denote the restored output images with the corresponding size. MSC loss is proposed in [14] and works similarly to the MSE loss. A small number ϵ is added to ensure that the result is not zero.

MSED loss is defined as

$$\mathcal{L}_{msed} = \sum_{k=1}^{3} \sqrt{\|\Delta(\hat{\mathbf{S}}_k) - \Delta(\mathbf{S}_k)\|^2 + \epsilon^2} \tag{3}$$

Δ denotes the Laplacian operation. The Laplacian operation extracts edges in the image and therefore MSED loss helps recover edges in the input images.

MSFR loss is defined as

$$\mathcal{L}_{msfr} = \sum_{k=1}^{3} \sqrt{\|\mathcal{F}(\hat{\mathbf{S}}_k) - \mathcal{F}(\mathbf{S}_k)\|_1} \tag{4}$$

$\mathcal{F}$ denotes the 2D Real FFT. The FFT divides images into their individual frequency components and therefore MSFR loss is specialized for recovering each frequency component.

The total loss function $\mathcal{L}$ is denoted as

$$\mathcal{L} = \mathcal{L}_{msc} + \lambda_1 \mathcal{L}_{msed} + \lambda_2 \mathcal{L}_{msfr} \tag{5}$$

We set λ_1 and λ_2 as 0.05 and 0.01, respectively, and ϵ as 0.001.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We use Qian *et al.* [1]'s dataset for training and evaluation of each method. The dataset is composed of pairs of images with and without raindrops by taking pictures of carefully chosen backgrounds using two glasses: one without any drop and one sprayed with water. The dataset has 861 pairs for training and 58 pairs for testing. While training, we augment the training dataset using the random horizontal flip and random crop to (256 × 256).

We set the number of the Res FFT-Conv blocks and GC blocks in each group (N and N_G) to 19 and 5, respectively. We decide these numbers by considering the balance between performance and computational cost. We set the batch size to 4 and train for 3,000 epochs. We set the initial learning rate to 0.0001 and decrease it by multiplying 0.5 every 500 epochs.

B. RESULTS

As for quantitative comparison, we use PSNR and SSIM, which are widely used to evaluate image restoration tasks including image raindrop removal. Table 1 shows the quantitative comparison of raindrop removal with several existing raindrop removal methods. As shown in the table, the proposed method is superior to any existing methods in PSNR. Additionally, the proposed method is superior to any existing methods except He *et al.* [3] in SSIM. Although He *et al.* [3] is superior to the proposed method in SSIM, the difference is less than 0.10 %. Thus, our proposed method achieves comparably better results in quantitative evaluation when compared to existing raindrop removal methods.

Fig. 4 shows the qualitative comparisons of each raindrop removal method. Images in the 1st and 2nd columns are the ground truth and input image, respectively, and the others are results from each raindrop removal method. Images in the 2nd and 4th rows are enlarged images of those in the 1st and 3rd rows within the red box, respectively. As shown in Fig. 4, unnatural patterns(e.g., whole images in the 2nd row, sky and large wall in the 4th row) are apparent in images of existing methods ([2] and [3]). On the other hand, our proposed method correctly removes raindrops and outputs images whose appearances are smooth and natural. Furthermore, the proposed method recovers the details (e.g., small windows in the 4th row) more realistically than existing methods.

TABLE 1. Quantitative comparison of raindrop removal with existing methods, the highest results are highlighted and the 2nd highest are underlined.

Method	PSNR(dB)	SSIM
Eigen et al. [18]	28.59	0.6726
Pix2Pix [31]	30.14	0.8299
Qian et al. [1]	31.57	0.9023
Quan et al. [2]	31.44	0.9263
He et al. [3]	31.33	**0.9297**
All in One [22]	31.12	0.9268
Ours	**31.76**	0.9288

TABLE 2. Ablation studies of raindrop removal, the highest results are highlighted and the 2nd highest are underlined. GC denotes all GC blocks in the proposed network, $\mathbf{Y}^{\mathbf{fft}}$ denotes modules which process feature maps in the frequency domain in Res FFT-Conv block, $\mathcal{L}_{msed}$ denotes the MSED loss (3), and $\mathcal{L}_{msfr}$ denotes the MSFR loss (4).

Method	PSNR(dB)	SSIM
MIMO-UNet+ [26]	31.11	0.9252
DeepRFT+ [5]	31.55	0.9278
w-o GC	31.48	0.9284
w-o $\mathbf{Y}^{\mathbf{fft}}$	31.18	0.9264
w-o $\mathcal{L}_{msed}, \mathcal{L}_{msfr}$	31.68	0.9272
Ours	**31.76**	**0.9288**

C. ABLATION STUDIES

We conduct ablation studies to show the effectiveness against the base methods for raindrop removal. The proposed method is compared with two methods, MIMO-UNet+ [26] and DeepRFT+ [5]. MIMO-UNet+ [26] uses residual blocks instead of Res FFT-Conv blocks in our model and does not have GC blocks. DeepRFT+ [5] uses Res FFT-Conv blocks and does not have GC blocks. Both methods are not proposed for raindrop removal, so we trained the two methods using the same dataset as IV-B. We set various learning environments as mentioned in each paper.

We also conduct studies to show the effectiveness of the three proposed components for better raindrop removal. The first is the introduction of GC blocks to capture long-range dependencies. The second is adding Res FFT-Conv blocks in the encoder and decoder blocks to process feature maps in the frequency domain. The third is the MSED loss and MSFR loss to better restore high-frequency components. We compared our proposed method with three models:

1) Removed all GC blocks from the proposed network (w-o GC)

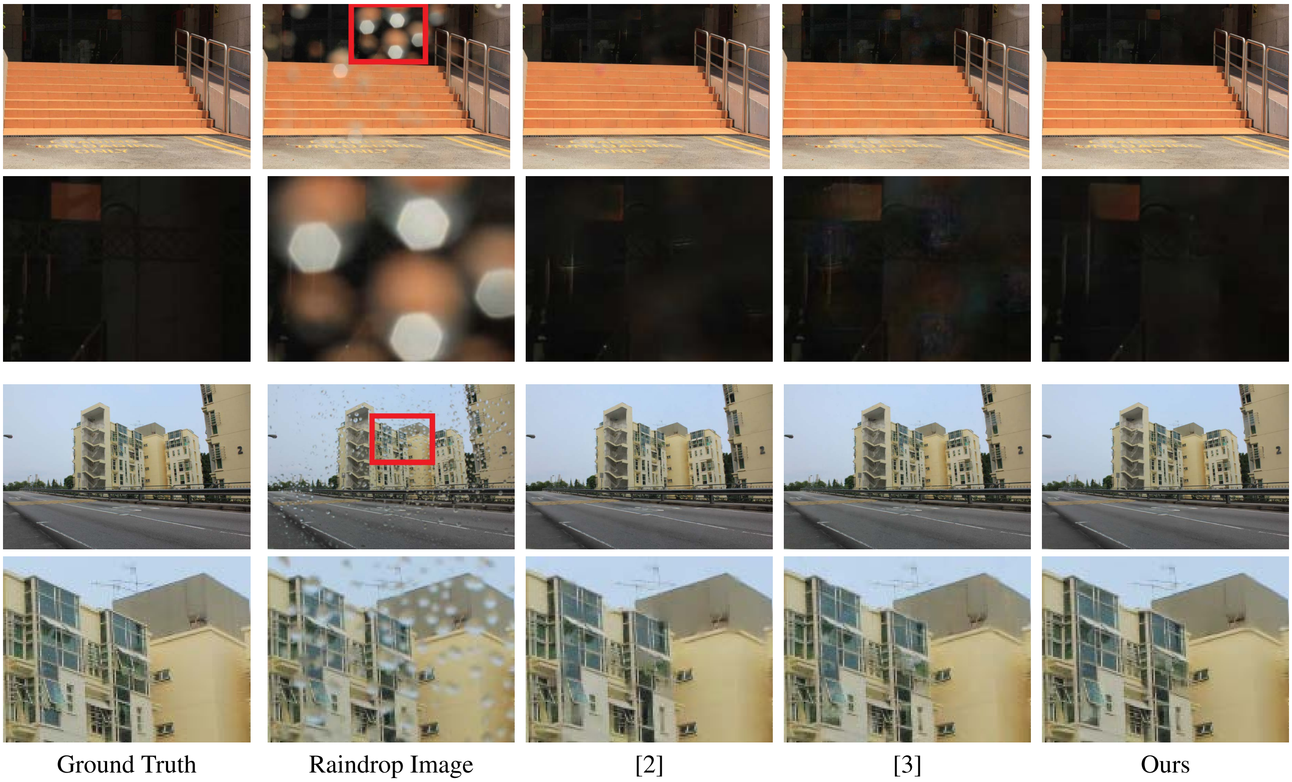


FIGURE 4. Qualitative results of each method on raindrop removal. Images in the 2nd and 4th rows are the enlarged images of those in the 1st and 3rd rows, respectively.

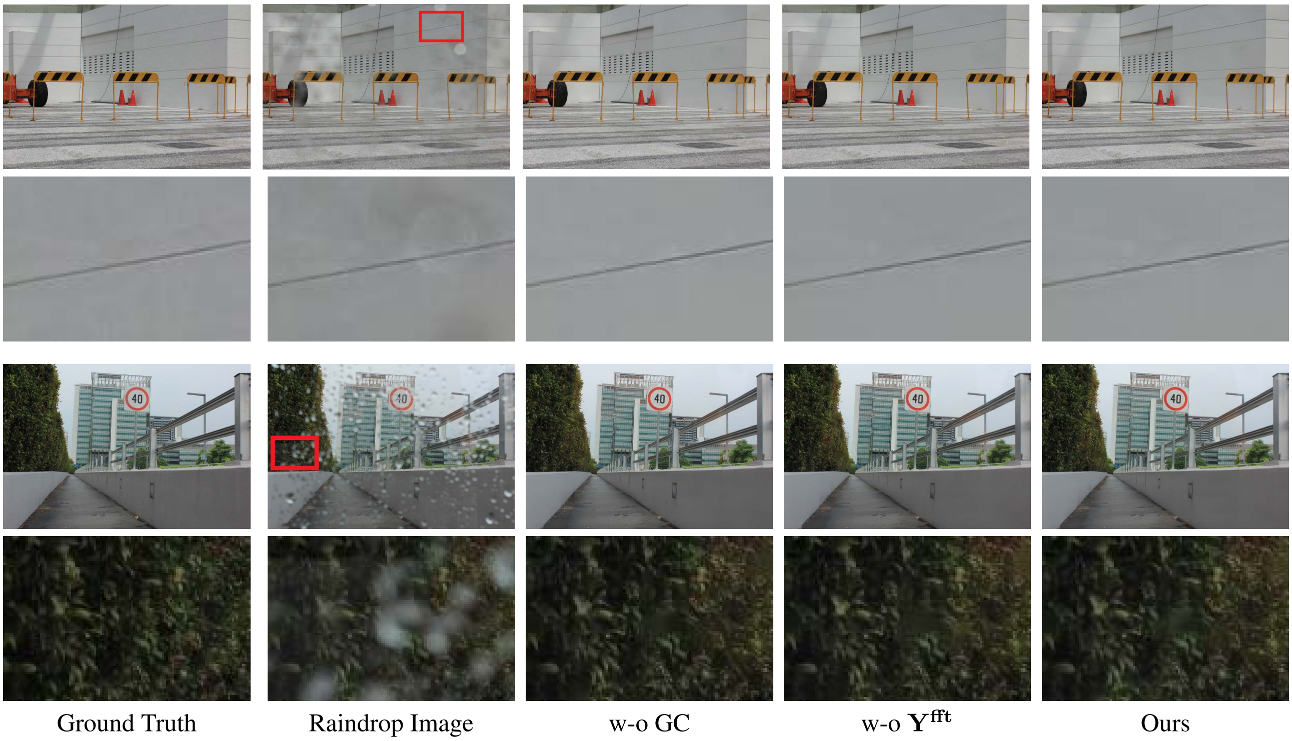


FIGURE 5. Qualitative results of each method on raindrop removal. Images in the 2nd and 4th rows are the enlarged images of those in the 1st and 3rd rows, respectively.

2) Replaced all Res FFT-Conv blocks in the proposed network with residual blocks (w-o $\mathbf{Y}^{\mathrm{fft}}$)

3) Removed the MSED loss and MSFR loss from the proposed total loss function (w-o $\mathcal{L}_{msed}$, $\mathcal{L}_{msfr}$)

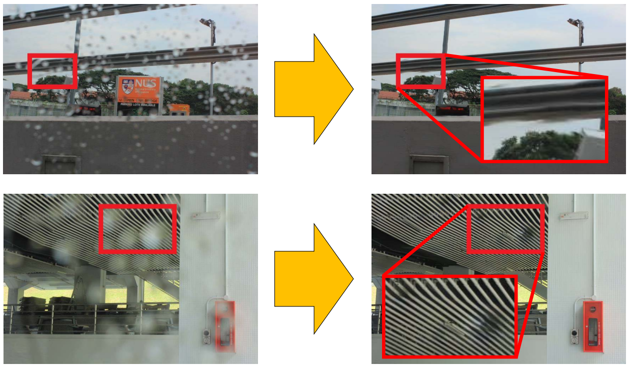


FIGURE 6. Examples of failures of raindrop removal in our proposed method.

In this study, we use the same learning environments in all models.

Table 2 shows the quantitative comparison of the 5 methods mentioned above. As shown in the table, our proposed network performs better in PSNR and SSIM than the two base methods. This shows that our improvements to the network structure, such as inserting the non-local operation, have quantitatively positive effect for raindrop removal. Table 2 also shows that our proposed network with all of the components performs the best in PSNR and SSIM compared to methods without each proposed component. This shows that each component has a positive effect on accurately removing raindrops from images and the introduction of processing feature maps in the frequency domain has the most significant contribution for better raindrop removal.

Fig. 5 shows the qualitative comparison of two methods (1. w-o $\mathbf{Y}^{\mathbf{fft}}$, 2. w-o GC) for raindrop removal. Images in the 1st and 2nd column are the ground truth and input image, respectively from Qian *et al.* [1]'s dataset, and the others are results from each raindrop removal method for ablation studies. Images in the 2nd and 4th row are enlarged images of those in the 1st and 3rd row within the red box, respectively. As shown in Fig. 5, though there are no significant differences between the three methods, using GC blocks tend to remove raindrops more clearly and generate more natural images since they are able to effectively capture features in distant locations. Such trend is also seen when we process the feature maps in the frequency domain.

D. FAILURES

Our proposed method shows excellent performance in raindrop removal both quantitatively and qualitatively, and handles various types of raindrops, from small and thin raindrops to large and dense raindrops. However, there are still some failure cases in our proposed model. For example, when more than one raindrop is connected and have an irregular shape (like the first row of Fig. 6). The propose model also fails when trying to inpaint fine patterns (like the second row of Fig. 6).

V. CONCLUSION

In this paper, we proposed a novel method for single image raindrop removal. Our proposed network utilizes feature maps in the frequency domain by introducing the architecture of Res FFT-Conv block. This component allows our network to recover high-frequency components more realistically. Additionally, our proposed network adds GC block to utilize features in distant places. This component allows our network to keep consistency throughout the output image. Experimental results show that these proposed components are effective in removing raindrops and that our proposed network achieves state-of-the-art performance.

REFERENCES

[1] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.

[2] Y. Quan, S. Deng, Y. Chen, and H. Ji, "Deep learning for seeing through window with raindrops," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2463–2471.

[3] D. He, X. Shang, and J. Luo, "Adherent mist and raindrop removal from a single image using attentive convolutional network," 2020, *arXiv:2009.01466*.

[4] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[5] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual Fourier transformation for single image deblurring," 2021, *arXiv:2111.11745*.

[6] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[7] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11908–11915.

[8] A. Singh, A. G. Bhave, and D. Prasad, "Single image dehazing for a variety of haze scenarios using back projected pyramid network," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Aug. 2020, pp. 166–181.

[9] T. Ye, M. Jiang, Y. Zhang, L. Chen, E. Chen, P. Chen, and Z. Lu, "Perceiving and modeling density is all you need for image dehazing," 2021, *arXiv:2111.09733*.

[10] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "DesnowNet: Context-aware deep network for snow removal," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3064–3073, Jun. 2018.

[11] W.-T. Chen, H. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 754–770.

[12] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I.-H. Chen, J.-J. Ding, and S.-Y. Kuo, "ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4176–4185.

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

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

[15] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "HINet: Half instance normalization network for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 182–192.

[16] M. Roser and A. Geiger, "Video-based raindrop detection for improved image registration," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, (ICCV) Workshops*, Sep. 2009, pp. 570–577.

[17] S. You, R. Tan, R. Kawakami, Y. Mukaigawa, and K. Ikeuchi, "Adherent raindrop modeling, detection and removal in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1721–1733, Oct. 2016.

[18] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 633–640.

[19] J. Ian Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, C. Aaron Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[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. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1–9.

[22] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3172–3182.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. New York, NY, USA: Springer, 2015, pp. 234–241.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] Y. Ganin and S. V. Lempitsky, "N^4-fields: Neural network nearest neighbor fields for image transforms," 2014, *arXiv:1406.6558*.

[26] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4641–4650.

[27] J. Cao, Y. Li, M. Sun, Y. Chen, D. Lischinski, D. Cohen-Or, B. Chen, and C. Tu, "DO-conv: Depthwise over-parameterized convolutional layer," 2020, *arXiv:2006.12030*.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[29] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. San Diego, CA, USA, Jun. 2005, pp. 60–65.

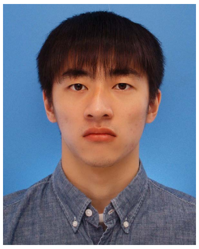
[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[32] C.-F.-R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[33] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "MDCN: Multi-scale dense cross network for image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2547–2561, Jul. 2021.

[34] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11447–11456.



SHINYA EZUMI received the B.E. degree in electronics and electrical engineering from Keio University, Yokohama, Japan, in 2022, where he is currently pursuing the M.E. degree under the supervision of Prof. Masaaki Ikehara. His research interest includes automatic image processing using deep learning, currently in crowd counting.

MASAAKI IKEHARA (Senior Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1984, 1986, and 1989, respectively. He was an Appointed Lecturer at Nagasaki University, Nagasaki, Japan, from 1989 to 1992. In 1992, he joined the Faculty of Engineering, Keio University. From 1996 to 1998, he was a Visiting Researcher at the University of Wisconsin–Madison, Madison, WI, USA, and Boston University, Boston, MA, USA. He is currently a Full Professor with the Department of Electronics and Electrical Engineering, Keio University. His research interests include multi-rate signal processing, wavelet image coding, and filter design problems.