Adherent Mist and Raindrop Removal from a Single Image Using Attentive Convolutional Network

Da He¹, Xiaoyu Shang¹, and Jiajia Luo²*

¹ University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, China
{da.he,shangxiaoyu}@sjtu.edu.cn

² Biomedical Engineering Department, Peking University, Beijing 100191, China
jiajia.luo@pku.edu.cn

Abstract. Temperature difference-induced mist adhered to the windshield, camera lens, etc. are often inhomogeneous and obscure, which can easily obstruct the vision and degrade the image severely. Together with adherent raindrops, they bring considerable challenges to various vision systems but without enough attention. Recent methods for similar problems typically use hand-crafted priors to generate spatial attention maps. In this work, we propose to visually remove the adherent mist and raindrop jointly from a single image using attentive convolutional neural networks. We apply classification activation map attention to our model to strengthen the spatial attention without hand-crafted priors. In addition, the smoothed dilated convolution is adopted to obtain a large receptive field without spatial information loss, and the dual attention module is utilized for efficiently selecting channels and spatial features. Our experiments show our method achieves state-of-the-art performance, and demonstrate that this underrated practical problem is critical to high-level vision scenes.

Keywords: Adherent mist, raindrops, removal, attention, deep learning

1 Introduction

High-level computer vision tasks, including object detection and segmentation, are highly dependent on the quality of captured images. However, outdoor vision systems such as security monitoring and vision-based self-driving cars or driving assistance systems are easily influenced by severe weather or other environments. As shown in Fig. 1(a), raindrops adhered to camera lens or windshield can highly degrade the image or obstruct the visibility.

Adherent raindrops (or waterdrops) exist widely in different scenes. Without enough shelter, rain streaks will result in raindrops naturally. In special environments like fishing-boats, splashing water can also form waterdrops. Raindrops

* Corresponding author.
Fig. 1. Examples of adherent raindrops and/or mist. (a) Raindrops adhered to a camera lens (Image source: The Nature Conservancy). (b) Inhomogeneous mist adhered to a windshield. (c) Mist and raindrops adhered to a camera lens

can have various visual effects because of various drop sizes, different distances to the camera, and complex refraction, which makes it extremely difficult to model the raindrops manually.

Besides raindrops, another severe and commonly-observed interference, adherent mist, is rarely mentioned in computer vision. As shown in Fig. 1(b), mist adhered to the windshield can also severely hamper the vision, which is vitally dangerous for driving.

Adherent mist is often caused by the temperature difference of air between two sides of the glass window. Unlike the haze in the air, mist adhered to glass windows is usually inhomogeneous (e.g., Fig. 1(b)).

Interestingly, adherent mists can sometimes develop along with the raindrops, especially for driving, where there is a temperature difference between the warm steam inside and the cooled glass by the raindrops outside. Therefore, it is necessary to handle both adherent mist and raindrops together. Fig. 1(c) shows the overlapping of mist and raindrops adhered to a camera lens.

By modifying the formulas in [3,18,20,23], the degraded image $I_{cap}$ captured by the camera can be modeled as a combination of raindrops $R$, the scattering of adherent mist $A$, and the clean background $B$ as follows:

$$I_{cap} = ((1 - M) \otimes B + R) \otimes t + A \otimes (1 - t)$$

where $M$ denotes to a binary mask showing the existence of raindrops, $\otimes$ means element-wise multiplication, and $t$ is the transmission map indicating how much information has passed through the adherent mist. $I_{cap}, M, B, R, t,$ and $A$ are location-related maps (i.e., matrices) instead of constants.

In this study, we aim to remove the adherent mist and raindrops from a single impaired image to obtain a clean image. The proposed approach, based on convolutional neural networks (CNNs), can automatically visually remove raindrops and mist of varying degrees. It can benefit high-level vision algorithms and improve their robustness in severe weather.

As both raindrops and mist might block some objects partially or even entirely, it is challenging to restore a purely clean image. To better solve the problem, we apply smoothed dilated convolution [27] into our model to obtain a relatively large receptive field instead of general pooling operations that might
lose spatial information. Dual attention modules consisting of channel attention and spatial attention are adopted for information selection and efficient learning. Perceptual loss is also applied in order to generate intuitively realistic images and avoid over-smoothing details [14,17].

Besides, to strengthen the spatial attention and handle the inhomogeneity, the classification activation map (CAM) [32] for each captured image is then calculated based on a newly trained classification network and acts as the spatial attention input concatenated with the original RGB image. The classification network is designed to accurately judge if the collected image is clean, with raindrops only, or with both raindrops and adherent mist. The proposed CAM attention can effectively locate important regions of the input image and improve the restoration performance, without hand-crafted priors.

Thus, our contributions are four-fold as follows:

- We address the visual interference problem caused by adherent mist and raindrops, which is frequent and essential but less studied. Without appropriate solutions, this problem possibly limits extensive outdoor camera-based systems.
- A CNN-based method is proposed to jointly remove adherent mist and raindrops. The proposed CNN is designed with smoothed dilated convolution, dual attention module, and perceptual loss to achieve state-of-the-art performance. The peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [28] values of the test set is improved from 17.9213 and 0.6283 to 21.8073 and 0.7549, compared with the degraded input.
- We introduce CAM attention as an efficient spatial attention input. Except for accessible classification labels, it does not require standard assumptions to find out worthy-attention pixel locations. This technique can be easily implemented in other low-level image generation problems.
- The improvements of high-level algorithms using enhanced images are demonstrated, showing the importance of both the problem and the proposed method. We also provide a new public dataset, which contains image pairs of ground truth (clean) images and degraded images (with raindrops and adherent mist). Further studies could use our dataset to develop new algorithms.

2 Related Work

Most visibility enhancement algorithms related to rain weather focused on rain streaks [7,18,19,26,31]. However, the shapes and physical effects of raindrops exceedingly differ from those of rain streaks. Thus, these streak removal algorithms can not be directly applied to raindrop removal.

A few studies about raindrop detection have been proposed for years. Kurihata et al. [16] utilize principal component analysis to learn the characteristics of raindrops, while Ito et al. [13] employ the maximally stable extremal regions to detect raindrop candidates. However, these methods do not concentrate on raindrop removal. You et al. [30] pay attention to both raindrop detection and
4 D. He et al.

removal in videos. They notice the temporal change of intensity of raindrop pixels is different from that of non-raindrop pixels, which is not applicable to the single image case. Eigen et al. [5] may be first to address the single image raindrop removal problem. They use a straightforward idea which trains a shallow CNN with pairs of raindrop images and clean images. Unfortunately, the limited capacity of the CNN model restricts its performance, especially for large and dense raindrops.

Qian et al.’s work (DeRaindrop) in 2018 [20] might be the first practical solution for the raindrop removal task. Specifically, they build a generative adversarial network (GAN) [8] to generate raindrop-free images. Besides the GAN, DeRaindrop also includes a recurrent network to gradually locate raindrops location in the input image, working as the spatial attention for the GAN. The performance of DeRaindrop is related to this raindrop mask, whose ground truth is calculated by the difference between the degraded image and the clean image with a hand-crafted thresholding 30. However, it is difficult to apply this manual threshold to images with adherent mist, which contain complex variations of pixel values. Quan et al. [22] propose shape-driven attention for raindrop regions, which is based on priors about the shape of raindrops.

Few works primarily focus on the adherent mist. Some similar studies mainly concentrate on atmospheric haze (or fog) [1,3,6,10,23]. On the one hand, they both have a scattering effect, which results in white color and degrades the original information. But compared to atmospheric haze, adherent mist is more likely to be inhomogeneous as it is related to the specific design of the equipment, such as which direction the warm steam comes from and how the window looks (e.g., mist adhered to the windshield near the drivers seat in Fig. 1(b)).

Single image dehazing is initially explored using some priors-based methods. Fattal et al. [6] work on estimating the albedo of the scene for single image dehazing. He et al. [10] propose a dark channel prior which talks about the local minimum of the dark channel varies between haze and haze-free images.

Then some learning-based methods have been proposed for dehazing using a single image. Cai et al. [1] introduce an end-to-end CNN to estimate the transmissions from images with haze. Ren et al. [23] propose multi-scale gated fusion network to improve the performance. And the smoothed dilation technique is adopted by Chen et al. [3] to remove gridding artifacts.

However, in our work, the superposition of adherent mist and raindrops results in complex local features. The influence of raindrops limits general dehazing algorithms using transmission estimation. It is also difficult to manually input an attention map similar to that in [20] or [22]. Therefore, the proposed network is designed to efficiently utilize features and restore images for the practical scene.

3 Methods

3.1 Network Architecture

The proposed network is shown in Fig. 2. The degraded image concatenated with its CAM attention map is used as the input for a generative network to output
Fig. 2. The architecture of the proposed network. The degraded input image is firstly used to generate a CAM attention map with a classification network. Then the input image concatenated with the CAM attention map is handled by the generative network to obtain the clean output. The details of the smoothed dilated convolution module and the dual attention module are elaborated in the following figure.

The clean image. The CAM attention is generated through a newly trained classification network. Since the classification network can accurately distinguish the type of the degraded image, it can also inform us which regions contribute more to the classification result. The generative network applies the architecture of an autoencoder with some shortcut connections. The input is initially convolved by two convolutional layers with dual attention modules. After a convolutional layer with stride 2, the spatial size of the features is decreased. Smoothed dilated convolution modules, as well as dual attention modules, are successively applied to handle those features further. Shortcut connections are inspired by the residual learning concept [11]. Finally, the feature maps are converted back to the original size by a transpose convolutional layer and tuned by two convolutional layers to generate the output. Instance normalization [25] is also applied in the generative network.

**CAM Attention** It might be helpful to handle the inhomogeneous features [20] as shown in Fig. 1 with additional spatial attention maps instead of pure CNNs.
Although related works (i.e., [20, 22]) have generated such attention maps, they use a manually set threshold or shape prior, which might be less feasible to do so for complex images with adherent mist and raindrops. Instead, we utilize CAM to adaptively generate the attention maps.

CAM is proposed by Zhou et al. [32], which contributes to the visual interpretation of a CNN “black box”. For a classification network, CAM can show the importance of various spatial regions in the image to the prediction result. This mechanism is achieved by weighting different channels of the final convolutional layer’s features (denoting as $f_k(x, y)$ for a spatial position $(x, y)$ of channel $k$) using the weights of the dense layer (denoting as $\omega_k$) with respect to the prediction result $t$, which can be described as [32]:

$$M_t(x, y) = \sum_k \omega_k f_k(x, y)$$

(2)

where $M_t$ denotes to the CAM result. The features are resized to the shape of the input before multiplication, and the activation map is generated by adding values from all channels for each spatial position and normalization.

The architecture of our simple classification network is briefly shown in Fig. 2, which is similar to the structure in [9], with 7 convolutional layers and 1 dense layer. A global maximum pooling layer is omitted in Fig. 2, which is used to flatten 2D features to 1D feature vector before the dense layer. A softmax activation function is applied to obtain the one-hot result.

Specifically, our classification task aims to distinguish images in three situations: clean, with raindrops (and without mist), and with both adherent mist and raindrops. Once the class of the image is accurately identified, CAM can hopefully explain which regions in the input image contribute more to the prediction. Four examples of CAM attention are shown in Fig. 3.

**Fig. 3.** Examples of CAM attention. The first row shows the original RGB images, and the corresponding CAM is shown below. (a–b) Images from [20] degraded by raindrop only. (c–d) with raindrops only. (c–d) Images collected in our dataset with both adherent mist and raindrops. All CAM results share the same color bar on the right.
For images with raindrops only (Fig. 3(a–b)), CAMs show high activation for raindrop regions. And for images with both adherent mist and raindrops (Fig. 3(c–d)), except for some highlighted raindrops, we can visualize some highlighted edge textures in CAMs (e.g., the tall building on the left of Fig. 3(c), and the grids on the wall in Fig. 3(d)). The attention results are reasonable as the CNN might classify the image based on the existence of raindrops and the edge variation behind the mist.

Raindrops regions and edge textures (i.e., high-frequency details) are actually the parts that we mostly focus on when observing the restoration results, which means regions that are important to the classification task are probably significant for the restoration task too. Therefore, we concatenate the original RGB image with its CAM and feed them to the generative network for image restoration.

Fig. 4. The architecture of the smoothed dilated convolution module and the dual attention module. (a) The smoothed dilated convolution module. The same module adopts the same dilation rate. The adding operation indicates the element-wise sum. (b) The dual attention module. The two dense layers in the channel attention part share the same weights for output from both GAP and GMP in each iteration. The multiplication operation in channel attention weights different channels, and the multiplication operation in spatial attention processes various spatial positions

**Smoothed Dilated Convolution Module** Dilated convolution is one of the useful ways to obtain a relatively large receptive field without decreasing the feature size like pooling. The convolution kernel covers small regions larger than the kernel size with some skipped pixels. However, dilated convolution might lead to gridding artifacts as described in [3]. To handle this disadvantage, [27] propose smoothed dilated convolution by adding a general convolution layer before the dilated convolution operation. As a result, the result of the dilated convolution is not only relied on spaced pixels but also some continuous information.

The smoothed dilated convolution module applied in our method is shown in Fig. 4(a). A standard convolution layer is added before each dilated convolution layer. When the dilation rate of the dilated convolution kernel is $d$, the kernel size of the corresponding standard convolution kernel should be $2d - 1$ in each dimension. In this way, the information from the skipped feature pixels in the
input feature map could also be used in the dilated convolution to avoid gridding artifacts. Specifically, there are 6 smoothed dilated convolution modules in our network as indicated in Fig. 2. The first 3 modules apply the dilation rate of 2, and the last 3 modules adopt the dilation rate of 4.

**Dual Attention Module** Considering the inhomogeneous distribution of mist and raindrops, another spatial attention mechanism with a self-attention concept besides the CAM attention is also adopted for feature selection. Besides, as introduced in [12], channel attention can also be useful to improve information extraction. These two attention parts form the dual attention module. And 12 modules are inserted in the generative network as indicated in Fig. 2.

As shown in Fig. 4(b), we build the dual attention module based on [29]. Firstly, feature maps are processed by global average pooling (GAP) and global maximum pooling (GMP), individually, and become 1D vectors with the length of the number of input channels. Then, the two vectors are fed into two shared dense layers individually before being added together. The shared dense layer means the same weights are used to the two inputs from GAP and GMP. The multiplication operation in the channel attention gives different channels in the feature tensor different weights according to the vector.

The spatial attention is followed by the channel attention. Maximum pooling along the channel axis (MPC) indicates that the maximum value at each spatial position is obtained by comparing pixels from all channels at the same position. Average pooling along the channel axis (APC) is a similar operation for average values. The multiplication operation weights different spatial positions in the input tensor according to the 1-channel attention map.

### 3.2 Loss Function

As indicated in [14,17], although pixel-wise mean squared error (MSE) might bring improvements on metrics like peak signal-to-noise (PSNR) and structural similarity index (SSIM) [28], the generated image is likely to be overly smoothed. Instead, a perceptual loss term can be applied to calculate the error from high-level views, which might be helpful in generating subjectively realistic images.

We calculate the perceptual loss based on the MSE of the output features from the 7th convolution layer of VGG19 [17,24]. The VGG19 model is pretrained on the ImageNet dataset [4], which might give a relatively high-level description of an image. Based on the restored image $I^{out}$ and the corresponding ground truth $I^{gt}$, the two output feature maps of the 7th convolutional layer are $f(I^{out})$ and $f(I^{gt})$, respectively. Then the perceptual loss is:

$$Loss_{per} = \frac{1}{n_x \times n_y \times n_c} \sum_{x=1}^{n_x} \sum_{y=1}^{n_y} \sum_{c=1}^{n_c} (f(I^{gt})_{x,y,c} - f(I^{out})_{x,y,c})^2$$  \hspace{1cm} (3)$$

where $x$, $y$, $c$ denote the dimensions of the feature map.
Finally, the perceptual loss is added together with the general MSE loss as:

$$Loss_{total} = \lambda_1 Loss_{mse} + \lambda_2 Loss_{per}$$

(4)

where $\lambda_1$, $\lambda_2$ are the weights of items and $Loss_{mse}$ is calculated by directly comparing $I_{out}$ and $I_{gt}$. $\lambda_1$, $\lambda_2$ was set as 1.0 and 0.05, respectively.

3.3 Dataset Preparation

To prepare the image pairs consisting of clean and degraded images, we take photos following strategies similar to those in [20]. Specifically, we utilize two same glass panels to simulate different cases. Firstly, we fixed the camera using a tripod. Secondly, one of the panels was randomly sprinkled with waterdrops on one side, and randomly sprayed with mist by a cosmetic sprayer on the other side. Then, we placed the blurred panel in front of the camera and took a photo as the degraded image. At last, the corresponding ground truth (i.e., clean) image was taken with the other clean glass panel.

The distributions of waterdrops and mist are random to keep good generalization. And the distance between the camera and the glass panel is also a random value between 0.5cm-4.0cm, thus the shape of waterdrops can have different appearances in the image due to the influence of camera focal length. The integrated camera of Redmi K20 Pro is used for taking pictures.

Finally, all the collected images were resized with bicubic interpolation to the shape of 640 × 480 to form the dataset. There are 572 image pairs for training, 73 image pairs for validation, and 72 image pairs for test. All the following evaluation results are based on images in the testing set.

4 Experiments

4.1 Experiment Setup

Both the classification network and the generative network are implemented on Keras framework with Tensorflow backend. Both models are trained using an Nvidia Titan RTX GPU with Adam optimizer. The learning rate of 1e-4 and the batch size of 8 were utilized for the classification network, while for the generative network, the learning rate is set as 1e-3, and the batch size is 4. During training, data augmentation operations, including flipping, rotation (180 degrees), brightness and contrast fluctuation, and gamma tuning, were applied.

In addition, only the collected dataset described in Sec. 3.3 is used to train the generative network. However, the image pairs of raindrops dataset from [20] are mixed with our dataset to train the classification network for raindrops classification. To avoiding data leaking, the testing images of our captured data that will be used to evaluate the performance of the image restoration are not used for training the classification network.

PSNR and SSIM are adopted for evaluating our model’s performance. Their values are calculated based on the luminance channel, so the RGB images are converted into YCrCb mode before evaluation.
4.2 Classification Results

The classification model was evaluated with 210 test images (66 from the dataset of [20], and 144 from our dataset). The classification network achieves the accuracy and f1-score of both about 96%. The relatively high classification performance might contribute to generate relatively reliable CAM attention results.

4.3 Restoration Results

To our knowledge, there are no other comparable studies that work on the exact same problem (i.e., jointly removal of adherent mist and raindrops). Thus, we compare the proposed method with the following three strategies:

(1) Our method is compared with two two-step combinations. A two-step combination method consists of a pre-trained general dehazing approach (that originally focuses on atmospheric haze) and the pre-trained raindrop removal model DeRaindrop (D) by [20] to restore the image. In other words, the degraded image is firstly processed by a dehazing model (i.e., GCANet [3] (G), or FFANet [21] (F)) and then by D. The two combinations are named as “G+D” and “F+D”, individually.

(2) We also tried to apply D before G (or F) for comparison. These two inverse combines are denoted as “D+G” and “D+F”, respectively.

(3) Although originally designed for other tasks, D, G, and F are advanced generative network that could be applied to our dataset. Therefore, we re-trained these three networks using our dataset. The three re-trained methods are denoted as “re-D”, “re-G”, and “re-F”, respectively.

The quantitative results of the three strategies above are listed in Table 1. DI stands for “Degraded Images”.

|          | DI     | G+D    | F+D    | D+G    | D+F    | re-D   | re-G   | re-F   | Ours   |
|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| PSNR     | 17.9213| 17.4249| 19.4054| 19.1926| 19.3927| 19.4788| 20.3022| 21.8073|        |
| SSIM     | 0.6283 | 0.6601 | 0.6825 | 0.6758 | 0.6831 | 0.6982 | 0.6844 | 0.7197 | 0.7549 |

The results of Table 1 show that, firstly, it is important to handle adherent mist and raindrops together. The co-existence of mist and raindrops might lead to horrible results that even worse than the degraded images (e.g., the PSNR result of G+D in Table 1). Secondly, the combination of F and D outperforms the combination of G and D with both processing orders. Thirdly, the three re-trained models generally show the effectiveness of re-training, especially for G and F. But our method always outperforms other methods.
In practical scenes, the combination of adherent mist and raindrops might severely impair the image, which might block objects from the subjective view. Examples in Fig. 5 and Fig. 6 exhibit the awful shooting environment and the restoration results from several methods above. Due to the space limit, we select D+G and F+D methods (Fig. 5), and re-F and re-D methods (Fig. 6) as representatives for visual comparison. The intuitive restoration examples of all the above methods are provided in the supplementary material.

From Fig. 5 and Fig. 6, we can find that some objects are not distinguishable in the degraded image. For example, in the first row of Fig. 5, it is hard to count how many cars in the degraded image, and in the second row of Fig. 6, the bicycles in front of the car almost disappear in the degraded image, which is vital for self-driving applications. If without the correct judgment, the safety of driving algorithms is to be doubted. This problem can be effectively handled by our method, as shown in the fifth column in Fig. 5 and Fig. 6. In addition, the proposed method can adaptively handle degradations with different extents. For both light degradation (e.g., the last example in Fig. 6) and severe degradation (e.g., the second example in Fig. 5), our method can always keep the reasonable brightness and fine textures.

According to the attached PSNR / SSIM values (Fig. 5 and Fig. 6), which are used to measure the similarity between the given image and the ground truth, it is proved that the proposed method can obtain relatively cleaner images and fewer artifacts than the other methods. The two combination methods (D+G and F+D in Fig. 5) sometimes result in serious luminance deviations. And the two re-trained methods (re-F and re-D in Fig. 6) show more artifacts than the proposed method.
Table 2. The ablation study results from different network designs

|                | Non-CAM | Non-CA | Non-SA | Non-SD | Ours  |
|----------------|---------|--------|--------|--------|-------|
| PSNR           | 21.2289 | 21.4883| 21.4294| 21.2753| 21.8073|
| SSIM           | 0.7494  | 0.7546 | 0.7514 | 0.7442 | 0.7549 |

The statistical results of these four models using our testing set are shown in Table 2. The performance decreases if any one of the above components is removed, especially for PSNR metrics. For example, without the CAM attention mechanism, the PSNR decrease by about 0.6dB. It is demonstrated that all those techniques effectively contribute to our model’s performance.
Fig. 7. Instance segmentation results using images from different sources: (From left to right) ground truth, degraded, F+D, re-F, and our method. Each row corresponds to the same sample. Small objects are best viewed by zoom-in

4.5 Applications

Besides the quantitative and qualitative evaluation in Sec. 4.2, the importance of the proposed problem and the effectiveness of our method are also validated in practical applications, such as for instance segmentation. We apply a cutting-edge segmentation algorithm, Cascade R-CNN method [2] from Detectron2 (https://github.com/facebookresearch/detectron2) model zoo to segment the images including the ground truth images, degraded images, and restored images as shown in Fig. 7. This advanced algorithm above also detects 2D objects with bounding boxes. For a fair and representative comparison, methods with relatively high-performance such as F+D and re-F are applied in this subsection. Segmentation results based on images restored by other methods are provided in the supplementary material.

As shown in the first row of Fig. 7 there is a car “missed” in the degraded image, the image from F+D method, and the image from re-F method. And in the second row, the segmentation algorithm can not find the bicycles in the degraded image, and F+D recovered image. Thus, it is proved that adherent mist degraded images not only impede our subjective views, but also impair the performance of computer algorithms. For the third sample, the segmentation algorithm mistakenly regards the building as a train in the degraded image, and F+D recovered image. The error possibly leads to severe consequences in practical applications.

In addition, panoptic segmentation is a newly proposed segmentation task [15] which tries to label all the pixels in the image. The results of panoptic segmentation based on various images are shown in Fig. 8. We applied R101-FPN method in Detectron2 (https://github.com/facebookresearch/detectron2) model zoo as the panoptic segmentation algorithm.
Fig. 8. Panoptic segmentation results using images from various sources: (From left to right) ground truth, degraded, F+D, re-F, and our method. For each scene, the segmentation results are attached below their corresponding images. The white bars in the second sample are added to hide the location information. Small objects are best viewed by zoom-in.

Panoptic segmentation might be more natural to show the influence of adherent mist and raindrops, as it demands all the spatial pixels of the image. In the first sample (1st and 2nd rows in Fig. 8), the segmentation algorithm cannot clearly find the border between the road and the grass, as well as the border between the tree and the sky in the degraded image. In the second sample (3rd and 4th rows in Fig. 8), segmentation results of the middle three images are horrible, especially for the degraded image. Because of the impact of adherent mist and raindrops, the trees, the building, and some bicycles cannot be identified by the panoptic segmentation algorithm. The segmentation result with the output from our method shows relatively distinguishable contours for the bicycles, the pavement, the trees, and the building.

5 Conclusion

In this work, we propose an image degradation problem caused by the co-existence of mist and raindrops adhered to a glass window or camera lens. This phenomenon shows different image characteristics compared to atmospheric haze or rain streaks. The problem is widely observed in reality but lacks investigation. Instead of using hand-crafted priors to generate spatial attention maps to handle this problem, we introduce CAM attention to generate attention maps
without priors, and utilize dual attention modules as well as smoothed dilated convolution modules in the restoration CNN to generate relatively clean images. Both quantitative evaluation and intuitive examples show quality improvement with the proposed method. The restoration with our method can also benefit high-level tasks such as instance segmentation and panoptic segmentation for generating reliable results, which demonstrates a vital significance in vision applications.

References

1. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing 25(11), 5187–5198 (2016)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
3. Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G.: Gated context aggregation network for image dehazing and deraining. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1375–1383 (2019)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
5. Eigen, D., Krishnan, D., Fergus, R.: Restoring an image taken through a window covered with dirt or rain. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 633–640 (2013)
6. Fattal, R.: Single image dehazing. ACM transactions on graphics (TOG) 27(3), 1–9 (2008)
7. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3855–3863 (2017)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
9. He, D., Cai, D., Zhou, J., Luo, J., Chen, S.L.: Restoration of out-of-focus fluorescence microscopy images using learning-based depth-variant deconvolution. IEEE Photonics Journal (2020)
10. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(12), 2341–2353 (2010)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
13. Ito, K., Noro, K., Aoki, T.: An adherent raindrop detection method using mser. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). pp. 105–109 (2015)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711 (2016)
15. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9404–9413 (2019)
16. Kurihata, H., Takahashi, T., Ide, I., Mekada, Y., Murase, H., Tamatsu, Y., Miyahara, T.: Rainy weather recognition from in-vehicle camera images for driver assistance. In: IEEE Proceedings. Intelligent Vehicles Symposium, 2005. pp. 205–210 (2005)
17. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4681–4690 (2017)
18. Li, S., Araujo, I.B., Ren, W., Wang, Z., Tokuda, E.K., Junior, R.H., Cesar-Junior, R., Zhang, J., Guo, X., Cao, X.: Single image deraining: A comprehensive benchmark analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3838–3847 (2019)
19. Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Single image rain streak decomposition using layer priors. IEEE Transactions on Image Processing 26(8), 3874–3885 (2017)
20. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2482–2491 (2018)
21. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. arXiv preprint arXiv:1911.07559 (2019)
22. Quan, Y., Deng, S., Chen, Y., Ji, H.: Deep learning for seeing through window with raindrops. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2463–2471 (2019)
23. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., Yang, M.H.: Gated fusion network for single image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3253–3261 (2018)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
26. Wang, G., Sun, C., Sowmya, A.: Erlnet: Entangled representation learning for single image de-raining. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5644–5652 (2019)
27. Wang, Z., Ji, S.: Smoothed dilated convolutions for improved dense prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2486–2495 (2018)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)
29. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19 (2018)
30. You, S., Tan, R.T., Kawakami, R., Mulaigawa, Y., Ikeuchi, K.: Adherent raindrop modeling, detection and removal in video. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(9), 1721–1733 (2015)
31. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 695–704 (2018)

32. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)
Supplementary Material

Introduction

The supplementary material contains three figures (in the following pages). As stated in the submitted article, the three figures show intuitive examples for all the compared methods mentioned in the manuscript. Part of the results has been exhibited in the main body.

According to the main text, some abbreviations are defined as follows:

- DI: Degraded Image
- G+D: GCANet \[3\] + DeRaindrop \[20\]
- F+D: FFANet \[21\] + DeRaindrop
- D+G: DeRaindrop + GCANet
- D+F: DeRaindrop + FFANet
- re-D: re-trained DeRaindrop
- re-G: re-trained GCANet
- re-F: re-trained FFANet

Specifically, Fig. S1 shows the restoration results of three samples. Fig. S2 demonstrates the instance segmentation performance based on images from different restoration operations. And Fig. S3 illustrates the panoptic segmentation masks using various restoration results.
### Restoration Examples

![Fig. S1](image) Examples of restoration results from different algorithms. Each row corresponds to the same sample. The values below images (except ground truth images) indicate the corresponding values of PSNR / SSIM.
Instance Segmentation Examples

Fig. S2. Instance segmentation examples using images from different restoration methods. Each row corresponds to the same sample. Small objects are best viewed by zoom-in. The black bars in some subfigures of the second sample are used to hide the location information.
Panoptic Segmentation Examples

Fig. S3. Panoptic segmentation examples using images from different restoration algorithms. For each scene, the segmentation results are attached below their corresponding images. Small objects are best viewed by zoom-in. The white bars in the ground truth of the second scene are used to hide the location information.