Page Index

Review #213534

Q1: There is no visualization for the two kinds of attention maps.

A1: We provide some visualization examples as shown in Fig. 1:

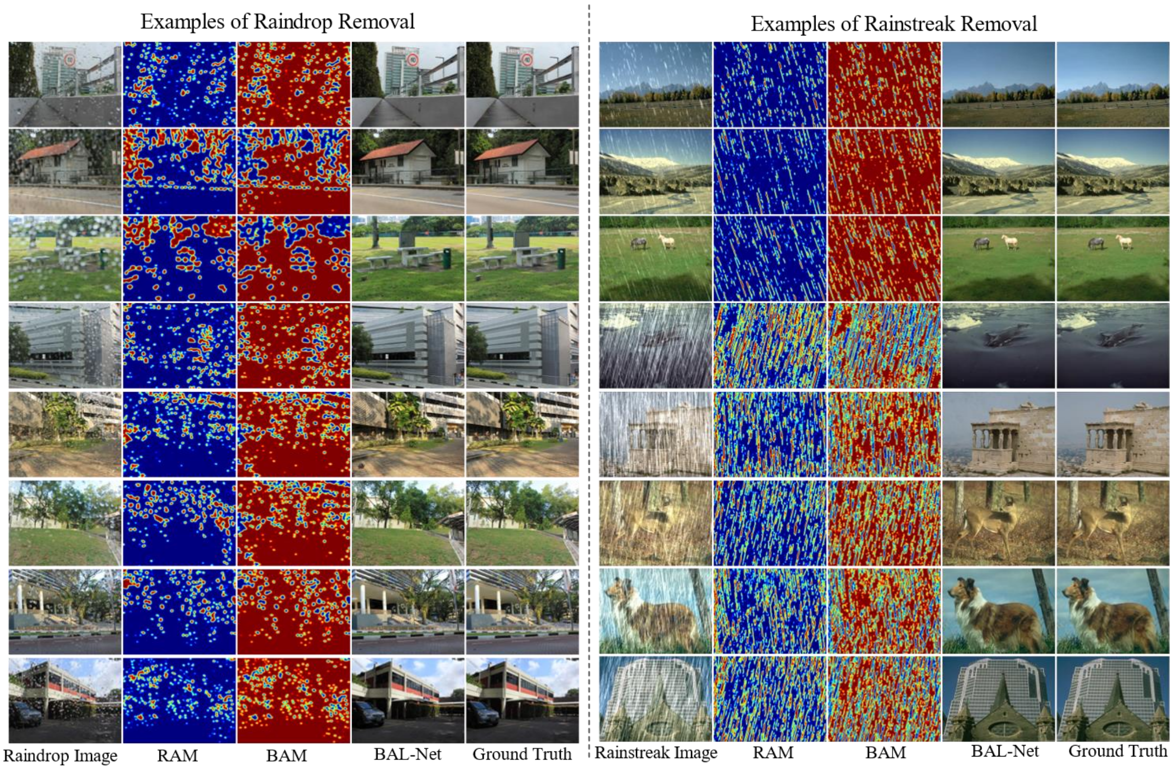


Figure1: Visualization of the generated bi-attention maps and the de-raining results (zoom in to see the results better).

Q2: Is it possible to just compute one attention map and take the opposite of it as the other attention map? How is it compared with the proposed bi-attention mechanism?

A2: We have done this experiment, and the specific network architecture for implementing this formulation is designed as Fig. 2:

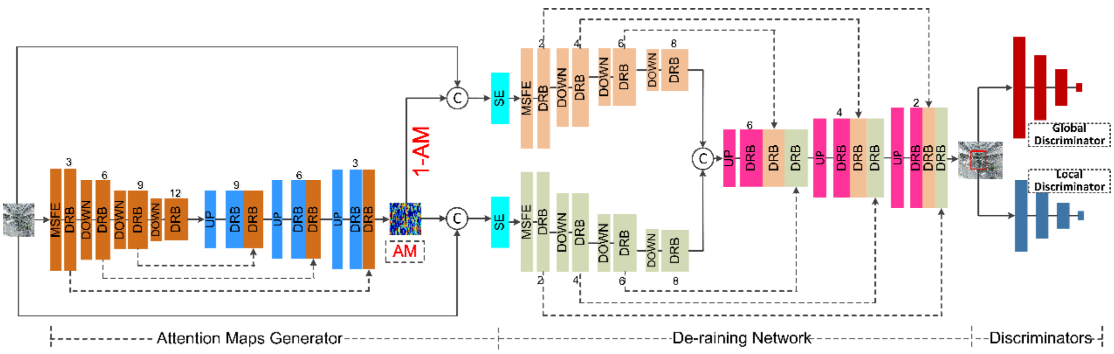


Figure2: Network architecture for utilizing single attention map and its opposite map for the bi-attention formulation. AM indicates attention map, and the red texts indicate the operation of generating the opposite map to obtain the bi-attention maps.

Following the network designed in Fig. 2, the new bi-attention formulation suggested by the reviewer is obtained. By defining rainy region attention map (RAM) or background attention map (BAM) as the choice of the AM, two different models are accordingly obtained. To evaluate the performance, as done in the paper, the AGAN data and the Rain100L data are selected for investigating the results on tasks of raindrop removal or rain streak removal. The corresponding de-raining results on AGAN data and Rain100L data are obtained as listed in Table 1:

Table 1: Quantitative results from models with different attention formulation

		Define AM as RAM	Define AM as BAM	Proposed
AGAN data	PSNR	32.20	32.26	32.48
	SSIM	0.9332	0.9335	0.9401
Rain100L data	PSNR	36.75	36.81	37.12
	SSIM	0.9592	0.9603	0.9758

As can be concluded from results shown in Table 1, the strategy suggested by the reviewer (compute one attention map and take the opposite of it as the other attention map) can obtain comparable results. However, the results are still worse than our proposed bi-attention mechanism, demonstrating the necessity of learning these two attention maps with different network branches specifically.

Q3: Ablation study on SE module to support its claim in paper.

A3: We have done this ablation study, and the specific results are shown in Table 2:

Table 2: Ablation study on the effect of SE module

SE module		×	√
AGAN data	PSNR	32.31	32.48
	SSIM	0.9368	0.9401
Rain100L data	PSNR	36.93	37.12
	SSIM	0.9705	0.9758

As can be seen in Table 2, the utilization of SE brings in improvements on both raindrop removal and rain streak removal tasks, demonstrating that the SE module is important for better taking advantage of the proposed bi-attention mechanism for the challenging single image de-raining task.

In the final version, we will combine the results in Table2 with the results in Sec4.2 of our submitted paper for providing a more complete ablation study.

Q4: What about the complexity of the model, such as parameter number, FLOPS, and running time?

A4: Both the model complexity and running time of the proposed model and other SOTA models are carefully calculated and provided in Table 3 and Table 4:

Table 3: Comparison with SOTA on model complexity and running time (s) on the 320*320 sized rainy image (results on rain streak removal).

	DSC	GMM	DDN	JORDER	DID	NLEDN	BAL-Net
Platform	CPU	CPU	GPU	GPU	GPU	GPU	GPU
#. Parameters	--	--	57,369	406,792	412,839	56,312,645	27,393,708
Running time (320*320)	118.4	421.8	0.19	375.4	0.21	1.65	0.62

Table 4: Comparison with SOTA on model complexity and running time (s) on the 320*320 sized rainy image (results on raindrop removal).

	Eigen	AGAN	BAL-Net
Platform	GPU	GPU	GPU
#. Parameters	26,427	36,206,378	27,393,708
Running time (320*320)	0.21	0.85	0.62

As can be seen from the comparison shown in Table 3 and Table 4, our model can provide a comparable running time and achieve new state-of-the-art de-raining results for both rain streak and raindrop removal task.

Review #260926

Q1: Lacks important formulation and mathematical modelling. In Equation (2), each term of loss function should be explained in mathematical formulations. The global discriminator and local discriminator should be illustrated in formulations.

A1: We have revised Figure 2 in the submitted paper for better illustrating how different loss is calculated, and the revised framework is shown in Fig. 3:

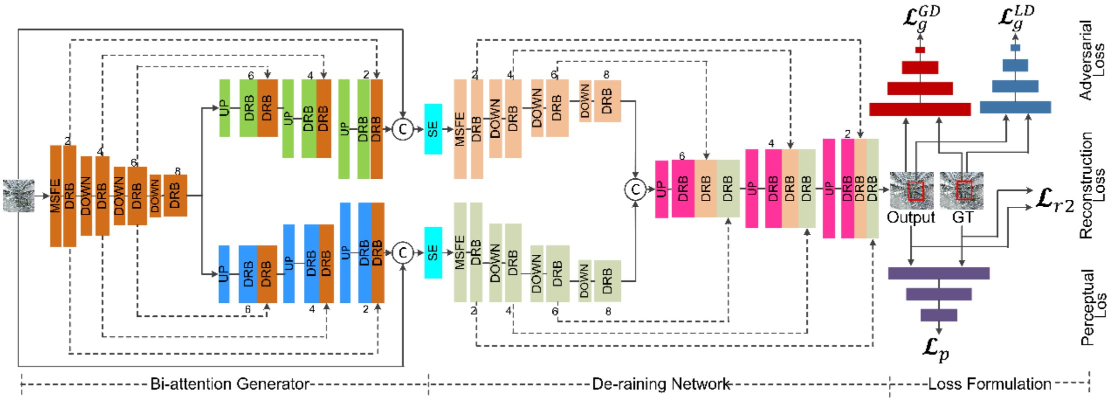


Figure3: Network architecture of the proposed BAL-Net (the specific loss formulation is illustrated in the rightmost part).

According to the illustration in Fig. 3, the specific formula for different losses can be defined as follows:

(1) Reconstruction Loss Formulation

The reconstruction loss is defined as the mean absolute error between the model output I_{output} and its corresponding groundtruth I_{GT}:

$$\mathcal{L}_{r2} = \left\| I_{output} - I_{GT} \right\|_1$$

Similarly, another reconstruction loss is defined in the same way, and used for training the bi-attention generation network:

$$\mathcal{L}_{r1} = \left\| AM_{output} - AM_{GT} \right\|_1$$

Where AM indicates the attention map.

(2) Adversarial Loss Formulation

In our implementation, we adopt the LSGAN for adversarial loss calculation, and the specific losses from the global and local discriminator are defined as follows:

$$\mathcal{L}_g^{GD} = \mathbb{E}\left[\left(D(I_{output}^{Global}) - 1\right)^2\right] \qquad \mathcal{L}_g^{LD} = \mathbb{E}\left[\left(D(I_{output}^{Local}) - 1\right)^2\right]$$

Where I_{output}^{Global} represents the images generated by the de-raining network, and I_{output}^{Local} represents randomly selected regions (70*70 in our paper) within the outputs from the de-raining network.

(3) Perceptual Loss Formulation

To improve the fidelity of the de-raining results, a pre-trained VGG network is also adopted to calculate the perceptual loss as follows:

$$\mathcal{L}_p = \frac{1}{CWH} \left\| F\left(I_{output}\right) - F\left(I_{GT}\right) \right\|_2^2$$

Where F represents features by a non-linear transformation with the pre-trained VGG-16, and we have assumed that the features are of size $\mathrm{W} \times \mathrm{H}$ with C channels. In our paper, we computer the perceptual loss from the layer relu2_2 of the pre-trained VGG-16 model. To train the overall network in an end-to-end manner, these losses are combined with suitable weight to form the overall loss as follows:

$$\mathcal{L}_{BAL-Net} = \lambda_{r1}\mathcal{L}_{r1} + \mathcal{L}_{r2} + \lambda_p\mathcal{L}_p + \lambda_g(\mathcal{L}_g^{GD} + \mathcal{L}_g^{LD})$$

Where we set $\lambda_{r1} = \lambda_g = 0.01$ and $\lambda_p = 0.05$ in our implementation.

Q2: In the theoretical part, the paper proposed the multi-scale architecture by observing that 'raindrops or rain streaks usually contaminate the background with different shapes, densities, and scales. The notations of the mentioned issues should be marked in the figure. Also, how the multi-scale technique can solve the issue should be better explained.

A2: First, more de-raining results with diversified rainy conditions are shown in Fig. 4:

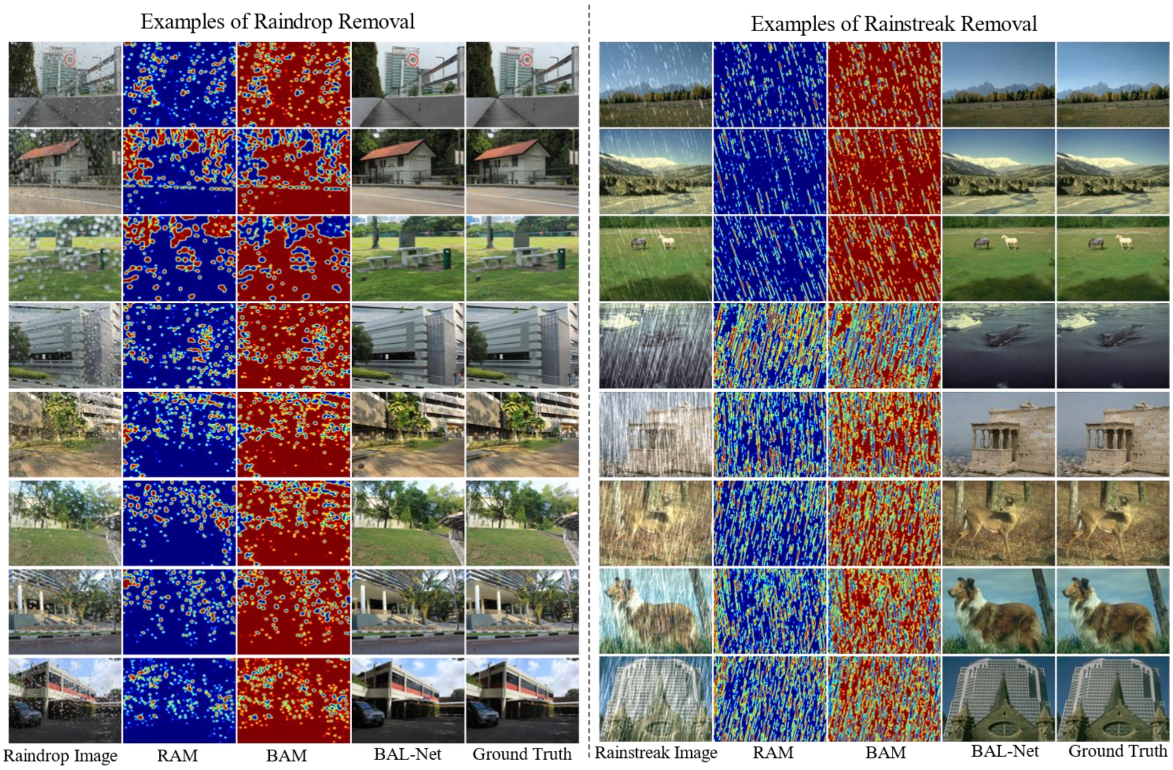
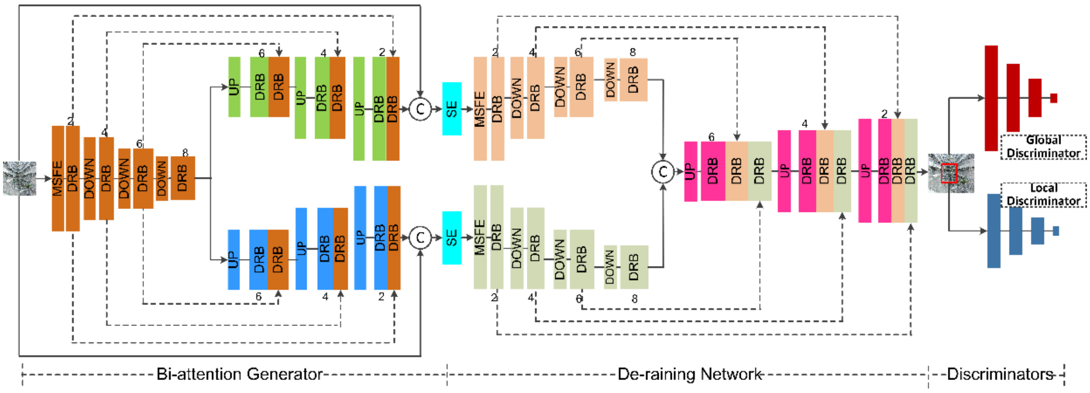
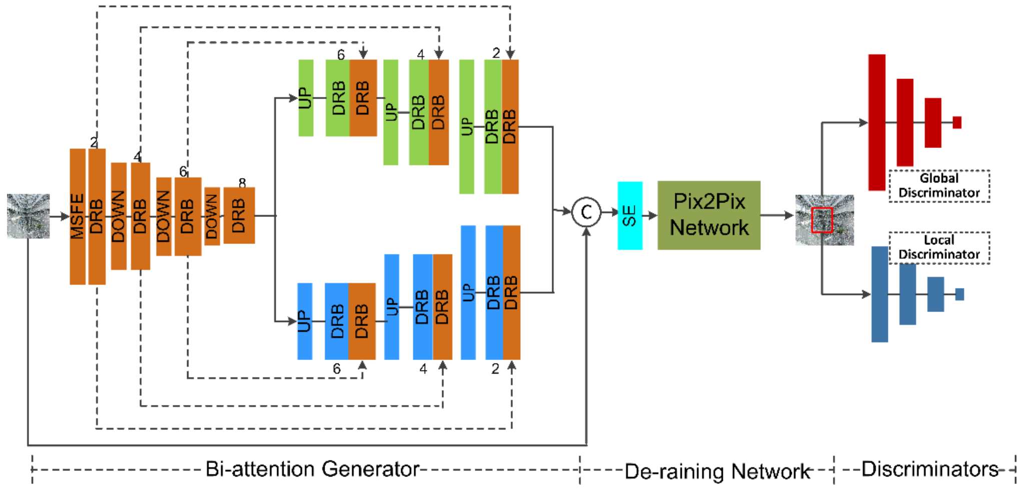


Figure 4: Results of BAL-Net with more diversified distribution of raindrops or rain streaks (zoom in to see the results better).

As shown in Fig. 4, the proposed BAL-Net can handle all these conditions well with extremely high-quality de-raining outputs. To further show the effect of the proposed multi-scale blocks, we have designed a comparative experiment to investigate the performance difference from two different setups: combine the proposed bi-attention mechanism with different de-raining network architecture including (1) the Pix2Pix network, and (2) the proposed multi-scale network. The network architecture for implementing these two setups are shown in Fig. 5:



(a) Setup of combining the bi-attention mechanism with the proposed multi-scale network



(b) Setup of combining the bi-attention mechanism with the lightweight Pix2Pix network

Figure 5: Network architecture of combining the proposed bi-attention mechanism with de-raining network designed as (a) complex network or (b) simple network.

With the network designed as Fig. 5(a) and Fig. 5(b), they are trained end-to-end with the loss $\mathcal{L}_{BAL-Net}$, and the specific de-raining results for both raindrop removal and rain streak removal are shown in Table 5:

Table 5: Ablation study on the effect of SE module

De-raining network design		Pix2Pix	Proposed Multi-scale Design
AGAN data	PSNR	31.94	32.48
	SSIM	0.9126	0.9401
Rain100L data	PSNR	33.98	37.12
	SSIM	0.8964	0.9758

As can be analyzed from Table 5, it is obvious that the de-raining results from network with multi-scale design are much better than the results from network without explicit multi-scale mechanism. Besides, for any other image-to-image translation tasks where Pix2Pix is used as the network design, it is possible to obtain better results by replacing the Pix2Pix network with the proposed multi-scale network.

Q3: In Figure 2, too much abbreviation words are used. I suggest building a notation table or explain the abbreviation in the caption. The visualized output of two bi-attention generation should be incorporated in the illustration.

A3: To improve the overall readability of the paper, we provide a table as follows to clearly explain the abbreviations used in our paper:

Table 6: Abbreviations explanation

Abbreviations	Full Names
BAL-Net	Bi-Attention Learning Network
UP	Up-convolutional layer
DOWN	Down-convolutional layer
SE	Squeeze-Excitation block
MSFE	Multi-scale Shallow Feature Extraction
DRB	Dense Residual Block
MMB	Multi-branch Multi-scale Block
MSC	Multi-Stream Convolution
DC	Dilated Convolution
SRM	Single image Restoration Model
CAM	Clean regions related Attention Map
RAM	Rainy regions related Attention Map

In the final version, we will provide this table for better reader reference.