Page Index

Review #213538

Q1: Explanation about Figure 2 is not very clear. Especially, why and how does the proposed method with soft-attention mask can address the imbalanced distribution issue.

A1: Intuitively, the motivation to our paper is the success of CNN in image inpainting: Given an image with some missing pixels, a CNN can be trained to learn knowledge from contextual regions to predict missing pixels well. The learned representation indeed depicts the image well as evidenced by ‘’Context Encoder’’. Similarly, raindrop removal refers to learning knowledge from the contextual regions to recover the pixels distorted by raindrops. Assuming that the representation learned by a raindrop removal network consists of two parts, one in charge of depicting the properties of the raindrop regions (Part-A) and another one for depicting the properties of contextual regions (Part-B). As in an image inpainting task, Part-B will provide information regarding the background to generate de-raining results while Part-A only provides information regarding raindrops distribution, serving as a guiding signal for the de-raining task. As a result, Part-B should play a much more important role for the de-raining task. However, due to the existence of the imbalanced distribution problem as illustrated in Figure 2, the quality of Part-B cannot match the one learned without the existence of raindrops pixels. Following this ‘inpainting idea’, we propose the idea of utilizing soft attention mask for dealing with the imbalanced problem, thus resulting in the enhanced context learning mechanism. Specifically, as shown in Figure 4, a two-branched encoder is designed, with one taking the raindrop image as input for learning representation including Part-A and Part-B and another branch taking the raindrop image together with the soft attention mask as input, and only learning with Part-B. By combining these two branches together, an enhanced Part-B is learned and much better results are obtained.

Q3: Authors are suggested to add visualization of computed soft-mask together with the raindrop removal results.

A3: We provide more results of the generated soft-mask and the corresponding de-raining results as shown in Fig. 1:

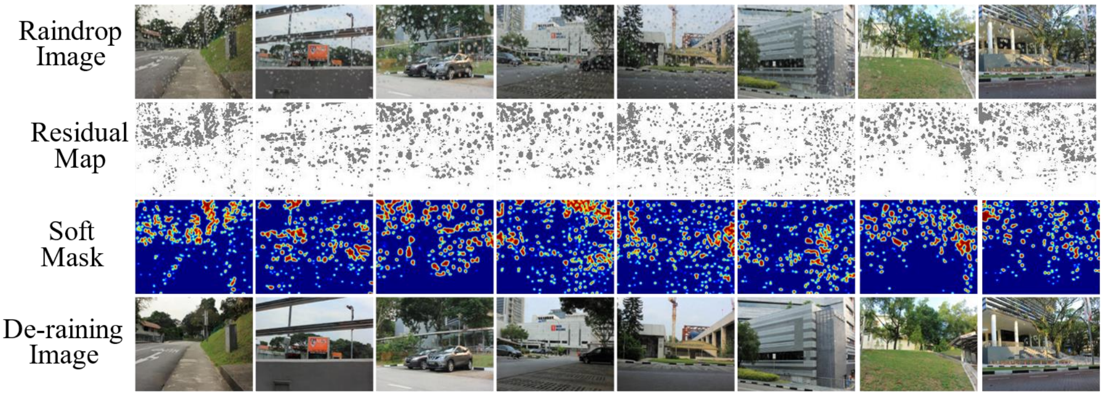


Figure 1: Visualization of the generated sot-mask and the corresponding de-raining results (zoom in to see the results better).

Q4: What is the speed of the proposed method at inference stage and how big is the model?

A4: Both the model complexity and running time of the proposed model and other SOTA models are carefully calculated and provided in Table 1:

Table 1: Comparison with SOTA on model complexity and running time (s) by processing a raindrop image with size of 320*320.

	Eigen	AGAN	EYE
Platform	GPU	GPU	GPU
#. Parameters	26,427	36,206,378	8,579,229
Running time (320*320)	0.21	0.85	0.39
PSNR	28.59	31.57	33.58
SSIM	0.6726	0.9023	0.9376

As can be seen from the comparison shown in Table 1, the proposed model can provide a comparable running time and achieve new state-of-the-art de-raining results for the challenging single-image raindrop removal task.

Review #252352

Q3: In table 2 and Fig. 7, EYE indeed outperforms previous methods. However, the complexity, e.g. parameter number, FLOPS, and running time, should be provided to demonstrate that, the gains come from the core idea instead of the increased parameter number.

A3: Both the model complexity and running time of the proposed model and other SOTA models are carefully calculated and provided in Table 2:

Table 2: Comparison with SOTA on model complexity and running time (s) by processing a raindrop image with size of 320*320.

	Eigen	AGAN	EYE
Platform	GPU	GPU	GPU
#. Parameters	26,427	36,206,378	8,579,229
Running time (320*320)	0.21	0.85	0.39
PSNR	28.59	31.57	33.58
SSIM	0.6726	0.9023	0.9376

As can be seen from the comparison shown in Table 2, the proposed model can provide a comparable running time and achieve new state-of-the-art de-raining results for the challenging single-image raindrop removal task.

Besides, to demonstrate that the performance gain mainly comes from the enhanced representation learning mechanism instead of the hierarchical multi-scale network design, as shown in Fig. 2, we propose to design the 'raindrop remover in Fig. 2' as the lightweight Pix2Pix network instead of the proposed network architecture, and observe whether the idea of soft-mask guided imbalance-aware representation learning mechanism can help improve the raindrop results of the Pix2Pix network. The specific network architecture for implementing such setup is shown in Fig. 2:

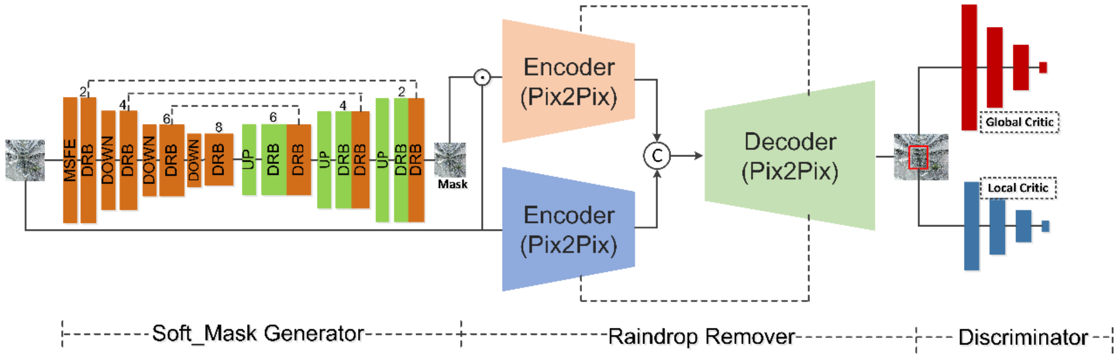


Figure 2: Network architecture of combining the proposed soft-mask with the lightweight Pix2Pix network as the design of raindrop remover.

With the network designed as shown in Fig. 2, the model is trained end-to-end to generate the de-rained images, and the specific raindrop removal results are shown in Table 3:

Table 3: Raindrop removal results obtained with different model setups

Models	Eigen	AGAN	Pix2Pix	Soft-Mask+Pix2Pix	EYE
PSNR	28.59	31.57	30.71	32.36	33.58
SSIM	0.6726	0.9023	0.8451	0.9218	0.9376

As can be observed from Table 3, after incorporating the proposed soft-mask mechanism into the Pix2Pix network, the performance is improved significantly, outperforming the existing state-of-the-art results by a large margin. Such results fully demonstrate that the core idea indeed work extremely well, even when incorporated into a simple network architecture without heavy parameters. Besides, results from the combination of soft-mask with Pix2Pix are much worse than the proposed EYE network which combines the soft-mask with an elaborately designed multi-scale architecture. Such results also demonstrate the effectiveness of the proposed multi-scale design in handling the diversified raindrops removal tasks.

Q4: In Fig. 4, the reviewer agrees that, the designed architecture can improve the capacity to percept context more globally. However, why does the design benefit mitigating the imbalanced distribution problem. Does the mask play an important role? If so, the attentive GAN has done this before.

A4: As you analyzed, the designed architecture is in charge of learning better contextual representation while the soft-attention mechanism plays an important role at mitigating the imbalanced distribution problem.

For the ‘’attentive GAN’’, we have experimented with their proposal (we did not include this analysis due the page limitation), and we found that the main improvement comes from injecting their attention maps to the discriminator, and this is similar with the conditional GAN formulation by injecting condition input to the discriminator for synthesizing better images. However, if the attention maps in attentive GAN are only injected to the generator, the improvement is very limited, demonstrating that their formulation has never considered and solved the proposed imbalanced distribution problem. Differently, our proposal solves these problems by designing a two-branched encoder network, and obtains far better results than the attentive GAN. Besides, it should be noted that we have also tried to inject our learned soft attention mask into the discriminator but did not observe any improvement, only causing training instability.

Q5: The architecture is complex and the ablation study is not comprehensive. Sec 4.2 is quite important. However, SR2, SR3, and SR4 are not explained. The reviewer cannot evaluate the merit of each part of the proposed method, including that of each module and each loss term.

A5: As indicated in the loss formulation described in the paper, there are altogether four losses are involved including the reconstruction loss, adversarial loss, and perceptual loss. In this answer, we firstly provide a new illustration of the propose framework to better demonstrate how each loss is formulated, and then give the detailed equations of each loss. Finally, we provide an ablation study on the effect of different loss terms.

The new framework is illustrated as follows:

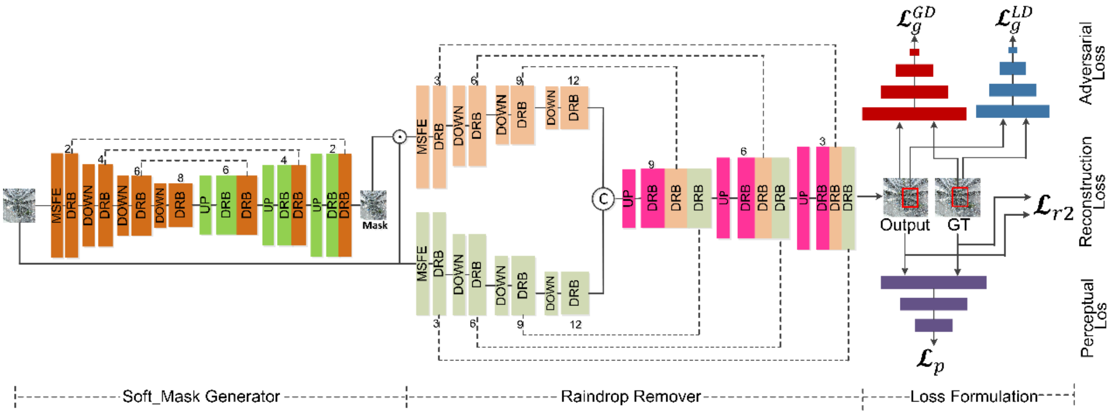


Figure 3: Network architecture of the proposed EYE (the specific loss formulation is illustrated in the rightmost part).

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 soft-mask generation network:

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

Where SM indicates the soft-mask.

(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.

With such formulation, we conduct the ablation study to investigate the effect of each loss. Specifically, we define the model trained with loss combining $\mathcal{L}_{r1}$ and $\mathcal{L}_{r2}$ as the basic model, on which the perceptual loss and adversarial loss are added subsequently. The experimental results are listed in Table 4:

Table 4: Ablation study on the effect of different loss terms

	Basic Model	Model_A	Model_B	Model_C	Model_D	Model_E
$\mathcal{L}_{r1}$	√	√	√	√	√	√
$\mathcal{L}_{r2}$	√	√	√	√	√	√
$\mathcal{L}_{p}$	×	√	×	×	×	√
$\mathcal{L}_{g}^{GD}$	×	×	√	×	√	√
$\mathcal{L}_{g}^{LD}$	×	×	×	√	√	√
PSNR	32.96	33.15	33.04	32.98	33.21	33.58
SSIM	0.9295	0.9324	0.9314	0.9301	0.9347	0.9376

As can be concluded from Table 4, the basic model trained with the reconstruction losses can obtain state-of-the-art results. On the basis of this, the incorporation of any other term can help improve the performance, and the combination of all these losses resulted in the best raindrop removal model.

In the final version, we will add these ablation study in Table 4 to Section 4.2 of the paper for providing a much more complete ablation study investigating the effect of each module and each loss term.

Review #260931

Q2: The loss function formula is not clearly written. What is the l1 loss? What is perceived loss? What is the global discriminator loss? What is the local discriminator loss?

A2: As indicated in the loss formulation described in the paper, there are altogether four losses are involved including the reconstruction loss, adversarial loss, and perceptual loss. In this answer, we firstly provide a new illustration of the propose framework to better demonstrate how each loss is formulated, and then give the detailed equations of each loss. Finally, we provide an ablation study on the effect of different loss terms.

The new framework is illustrated as follows:

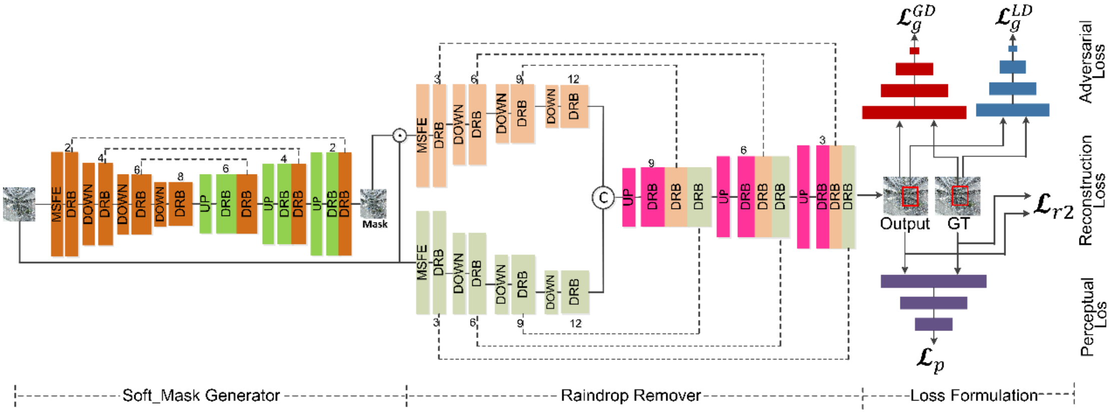


Figure 4: Network architecture of the proposed EYE (the specific loss formulation is illustrated in the rightmost part).

According to the illustration in Fig. 4, the specific formula for different losses can be defined as follows:

(4) 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 soft-mask generation network:

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

Where SM indicates the soft-mask.

(5) 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:

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

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.

(6) 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.

With such formulation, we conduct the ablation study to investigate the effect of each loss. Specifically, we define the model trained with loss combining $\mathcal{L}_{r1}$ and $\mathcal{L}_{r2}$ as the basic model, on which the perceptual loss and adversarial loss are added subsequently. The experimental results are listed in Table 5:

Table 5: Ablation study on the effect of different loss terms

	Basic Model	Model_A	Model_B	Model_C	Model_D	Model_E
$\mathcal{L}_{r1}$	√	√	√	√	√	√
$\mathcal{L}_{r2}$	√	√	√	√	√	√
$\mathcal{L}_p$	×	√	×	×	×	√
$\mathcal{L}_g^{GD}$	×	×	√	×	√	√
$\mathcal{L}_g^{LD}$	×	×	×	√	√	√
PSNR	32.96	33.15	33.04	32.98	33.21	33.58
SSIM	0.9295	0.9324	0.9314	0.9301	0.9347	0.9376

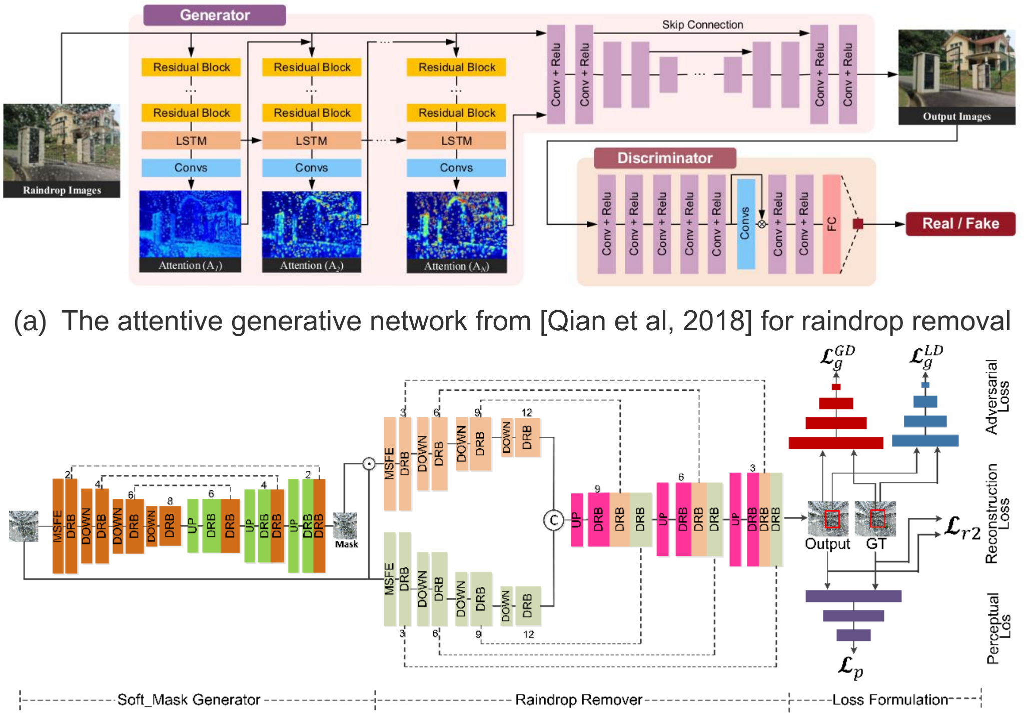
As can be concluded from Table 5, the basic model trained with the reconstruction losses can obtain state-of-the-art results. On the basis of this, the incorporation of any other term can help improve the performance, and the combination of all these losses resulted in the best raindrop removal model.

In the final version, we will add these ablation study in Table 5 to Section 4.2 of the paper for providing a much more complete ablation study investigating the effect of each module and each loss term.

Review #260932

Q1: What is the benefit of the proposed neural network compared with [Qian et al, 2018] from neural network structure's perspective? It seems to be not very clear.

A1: Specifically, the framework designed in [Qian et al, 2018] and the proposed EYE are shown as follows in Fig. 5(a) and Fig. 5(b):



(a) The attentive generative network from [Qian et al, 2018] for raindrop removal

(b) The proposed EYE network for raindrop removal

Figure 5: Illustration of two representative raindrop removal frameworks.

As shown in Fig. 5(a) and Fig. 5(b), both the attentive generative network (AGAN) from [Qian et al, 2018] and the proposed EYE network share a similar configuration in the main network parts, consisting of two parts in terms of attention generator and the de-raining network. Based on this observation, the difference in network structure covers the specific difference in the attention generator part, the de-raining network part, or both. To get insight on the benefit of the network structure, two different experimental setups are considered as follows:

Setup-1: Investigating the benefit of attention generator structure:

For this experiment, we fix the structure of the de-raining network, and then define the structure of the attention generator as the one from AGAN or EYE. The corresponding comparative results are shown in Table 6:

Table 6: Results on investigating the benefit of attention generator structure

Group1 (Fix the structure of de-raining network as the one from AGAN)				Group2 (Fix the structure of de-raining network as the one from EYE)		
	Attention generator structure				Attention generator structure	
AGAN	√	×		AGAN	×	√
EYE	×	√		EYE	√	×
PSNR	31.57	31.78		PSNR	33.58	32.89
SSIM	0.9023	0.9081		SSIM	0.9376	0.9294

As can be observed from the comparative results in Table 6, regardless of the structure of the de-raining network, better raindrop removal results are obtained by defining the structure of the attention generator as the one from EYE instead of that from the AGAN. Such results demonstrating the benefit of the proposed attention map generator structure.

Setup-2: Investigating the benefit of de-raining network structure:
For this experiment, we fix the structure of the attention generator, and then define the structure of the de-raining network as the one from AGAN or EYE. The corresponding comparative results are shown in Table 7:

Table 7: Results on investigating the benefit of de-raining network structure

Group1 (Fix the structure of attention generator as the one from AGAN)				Group2 (Fix the structure of attention generator as the one from EYE)		
	De-raining network structure				De-raining network structure	
AGAN	√	×		AGAN	×	√
EYE	×	√		EYE	√	×
PSNR	31.57	32.89		PSNR	33.58	31.78
SSIM	0.9023	0.9294		SSIM	0.9376	0.9081

As can be observed from the comparative results in Table 7, conclusion that is similar to that drawn from Table 6 is obtained as follows: regardless of the structure of the attention generator, better raindrop removal results are obtained by defining the structure of the de-raining network as the one from EYE instead of that from the AGAN. Such results demonstrating the benefit of the proposed de-raining network structure

As demonstrated by the results from Table 6 and Table 7, significant improvement are observed by utilizing our structure design as the design of the attention generator or the de-raining network. Such improvement fully demonstrates the superiority of our proposal over the one in [Qian et al, 2018] from neural network structure's perspective.
In the final version, we will also add this result to Section 4.4, thus providing a more complete comparison with state-of-the-arts.