1.1 Motivation and Contribution:

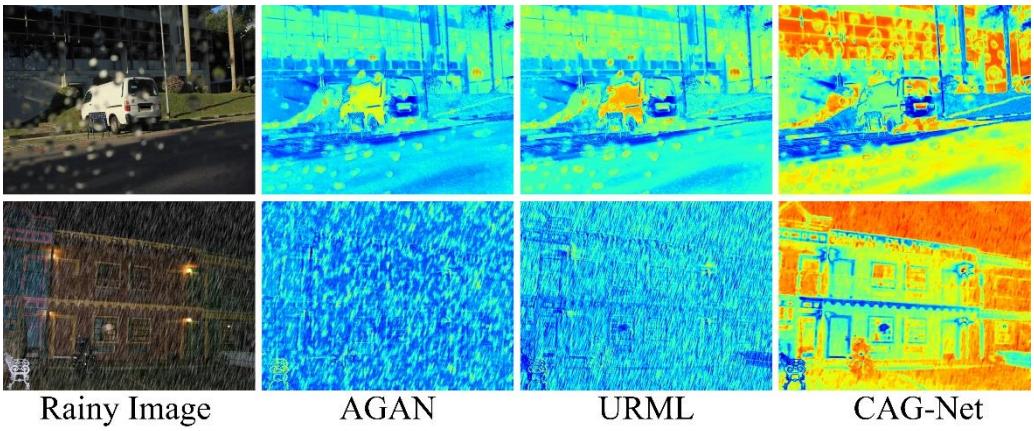


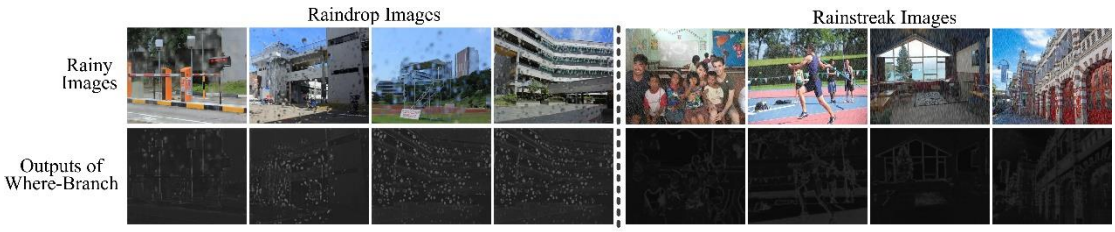
Fig. R1 Selected filter activation map from different prior guided de-raining models.

Existing DCNN-based de-raining models can be roughly categorized into two groups: Prior-Guided (PG) models and direct Rainy-to-Clean (R2C) translation models. The PG models are said [1-2] to utilize the priors (e.g., density label and rain location) to guide the filters of the de-raining network to focus more on the rainy regions, thus producing much better de-raining results than the R2C models. However, by delving deep into the filter behavior of two representative PG models, AGAN [1] and URML [2], it was found that the activations of feature maps, as shown in Fig.R1, are not well located at the rainy regions and the valuable contextual regions, thus producing unsatisfactory results with severe detail missing (as analyzed in Fig.1 of the paper). Both the unfavorable de-raining results and filter behavior indicate the deficiency of existing PG designs by simply concatenating the prior information with the rainy images. To improve this and fully explore the benefits of prior information, an innovative coarse-to-fine location prior utilization strategy is devised and used to formulate a three-stage de-raining network, which is demonstrated to be able to (1) deal with different rainy conditions (e.g., different density, shape, orientation), (2) generalize well to unseen examples, and (3) most importantly perform well on real rainy images. Besides, the proposed model can be generally used to solve many other problems: (1) the proposed prior utilization module can also be seamlessly incorporated into most of existing PG models for improving their performance (as demonstrated by results in Sec 4.5 of the paper); and without any modification, the overall framework has been successfully used for (2) producing SOTA result for the image denoising task (as demonstrated by the results in Table 4 of the paper) and (3) solving the landmark guided face manipulation problem and the keypoint guided pose transfer problem (results will be provided in the camera ready version).

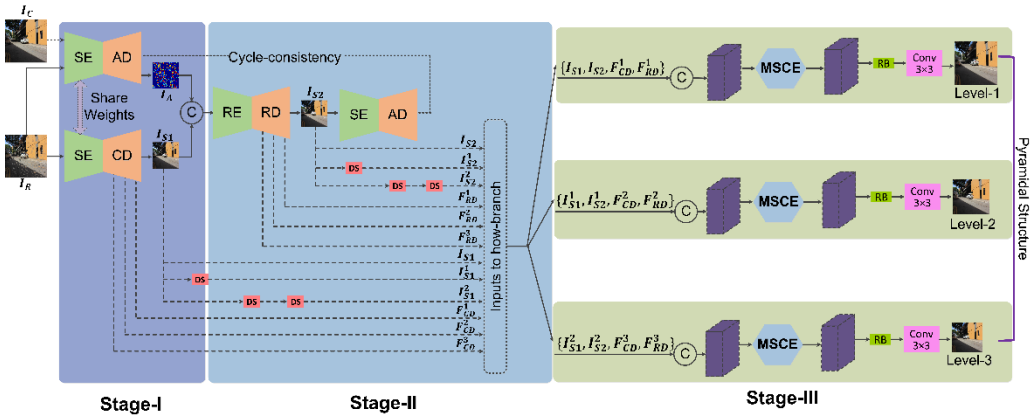
Relationship to existing PG models: Only the location map concatenation setup in stage-I share slightly similar design with existing PG models. However, the proposed coarse-to-fine prior utilization strategy consists of three stages, and two most important contributions are the attention-consistency formulation in stage-II and the flexible “where-and-how” module in stage-III.

1.2 Ablation study on effect of “where” branch in Stage-III

By taking attention maps and features from previous stages as input, the where branch is able to produce some weight maps (as shown in Fig.R2(a)), which can filter out the unfavorable regions while preserving the good regions on the outputs of how-branch in an adaptive manner, thus producing favorable de-raining results without over/under-deraining artefacts. By removing the where-branch from the complete model (the architecture after removing where-branch is shown in Fig. R2(b)), the PSNR/SSIM values drop sharply as in Table.R1, demonstrating the significance of the where-branch.



(a)



(b)

Fig.R2 (a) Examples of outputs from where-branch, (b) network architecture after removing where-branch. See here for higher resolution images.

Table R1: Average PSNR/SSIM values w/wo the Where-Branch.

Where-Branch	RS-Data (Test1)	RD-Data (TestA)
√	30.79/0.932	31.92/0.935
×	28.42/0.918	30.57/0.920

Results in Table.R1 will be added to Table 1 of the paper in the camera-ready version.

1.3 More results and analysis on real rainy images

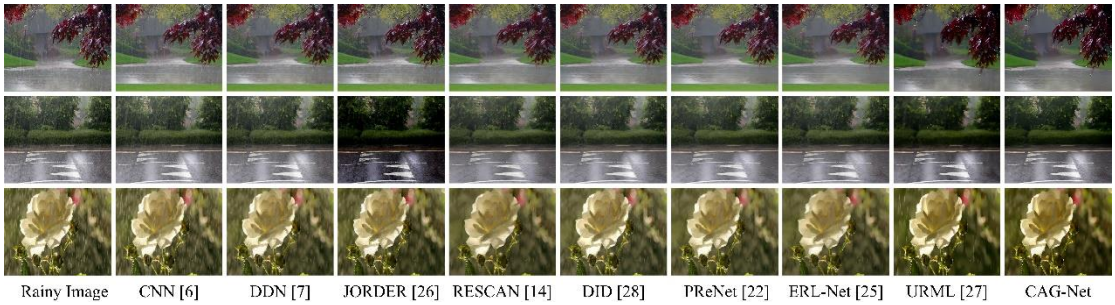


Fig. R3. Visual quality comparison on some real rainy images (click here to see more)

As shown in Fig. R3, very promising results are obtained by the proposed CAG-Net, which

can on the one hand remove the rain streaks thoroughly, and on the other hand recover the detailed structures well with high contrast. On the contrary, the other methods usually generate very blurry results with important details missing. For example, compared with our CAG-Net, the results of JORDER and URML show over-deraining results and some important structural details of background cannot be preserved/recovered well (e.g., the result in the 2nd row is very dark with low contrast, and the structure of the wall in the 1st row is not recovered well).

1.4 Loss formulation in Stage-III

Eq. (11) in the paper will be revised as follows by including λ_1 and λ_2 as the weights for the 2nd and 3rd pyramidal levels.

$$L_{S3}^{DR} = \left[\frac{I_{S3} - I_C}{U_4} + \log U_4\right] + \sum_{s=1}^{2} \|I_{S3}^s - I_C^s\|_1$$

2.1 Response to specific comments (#1):

Q1: As described in Line488, the term "gate convolution" is used to describe the operation by "where-branch". Other places that describe the where-branch as attention related operation will be revised in the camera-ready version.
Q2: "ME" in Fig.2 will be revised to "SE" in the camera-ready version.

2.2 Response to specific comments (#2):

Q1: The term "cycle-consistency" will be revised to "attention-consistency". The cycle-consistency constraint is described as high-level semantic loss under the assumption of defining the attention generation process as a binary segmentation task. Furthermore, as you will see in the code page released after paper acceptance, the learned attention map not only indicates the location but also provides information of rain density, direction, and shape etc. By this way, we follow this paper [6] to define the constraint from such a task as high-level semantic loss.
Q2: The "where-and-how" learning mechanism is a novel parallel feature refinement design, not a knowledge distillation process. We will revise this description in the camera-ready version.
Q3: See Sec1.1 for a clearer explanation on the novelty and contribution of the proposed model. The uncertainty map in our paper is different from others because (1) The network design for learning the uncertainty map (UM) is different from the others, and the UM is learned from a specified location-aware subnetwork in our paper. Furthermore, we have tried to replace UM as other designs in our model, and the PSNR value drops more than 0.6dB on both RS-Data and RD-Data. (2) More experiments after the paper submission provide a finding that UM only helps a faster network convergence, and the model without UM is able to achieve the same quantitative results with a larger training epoch (e.g., additional 200 epochs for RD-Data and additional 40 epochs for RS-Data). We will add this

finding in the camera-ready version.

Q4: The description on the learning of location detection and rain removal as a multi-task learning setup is stressed for supporting the lightweight shared-encoder design. However, the description of "multi-task" is not proper and we will just call it as "joint learning of location and rain removal" for easier understanding. On the other hand, the location detection sub-network in our design is highly different from the one in JORDER and is significant because it is used for (1) constructing the novel regularization term (attention-consistency constraint) in Stage-II and (2) learning the where-branch in Stage-III. Both designs have been demonstrated to improve the deraining results by a very large margin as evidenced by the elaborate ablation study. Also see Sec 1.1 on illustrating the contribution of the overall network design.

Q5: Results from some other SOTA models have been obtained by using the pre-trained model provided by the authors, and the quantitative results are as follows:

Table R2: Average PSNR/SSIM values on synthetic rainstreak datasets.

		SPA-Net [3]	Heavy-Net [4]	DAF-Net [5]
RS-Data (Test1)	PSNR	28.64	25.42	29.08
	SSIM	0.913	0.813	0.920
RS-Data (Test2)	PSNR	25.18	24.01	26.57
	SSIM	0.872	0.786	0.921

Results from all these additional models are much worse than our CAG-Net, and results in Table.R2 will be added to Table 2 of the paper in the camera-ready version.

Q6: The influence of the number N in Eq. (6) is investigated with a similar model setup as described in Sec4.3, and the results are in Table. R3. It can be observed that the de-raining results will be stable after N=15. Consequently, to achieve a good tradeoff between performance and inference speed, N is set as 15 in all the experiments.

Table R3: Average PSNR/SSIM values with different N.

Value of "N"	RS-Data (Test1)	RD-Data (TestA)
1	28.46/0.920	30.61/0.921
5	29.67/0.924	30.89/0.925
10	30.47/0.929	31.59/0.930
15	30.79/0.932	31.92/0.935
30	30.74/0.932	31.91/0.936
48	30.64/0.926	31.84/0.931

Results in Table.R3 and corresponding analysis will be added to Sec4.3 of the paper in the camera-ready version.

Q7: Reference number for DID in Fig. 6 and Fig. 8 should be [28], and we will correct this in the camera-ready version.

Q8: In line 442-443, we explain why only one level is described, and the final result is the one corresponding to the full-size output, which refers to I_s3 in the paper.

Q9: We simply follow AGAN to select the threshold for binarizing the rainy layer. Besides, we have also tried other values and the average PSNR results for both RS-Data and RD-Data are worse than the mean pixel value.

2.3 Response to specific comments (#3):

Q1: As introduced in Sec1.1 of the letter, existing de-raining models can be categorized into two groups: PG models and R2C models. By re-implementing most SOTA models of each group (our implementation of these models will be released in the GitHub page after paper acceptance), we found that PG models generally perform much better than the R2C models with the prior guidance, especially the location prior guidance. Subsequently, driven by the curiosity on how the PG models work, we delve deep into the mechanism of existing models and occasionally find their deficiency by analyzing the filter behaviors. Motivated by this, we propose to overcome the deficiency and design a more advanced prior utilization framework, and finally derive the CAG-Net. The explanation on the formulation of each module and stage are carefully provided at each Section, and also elaborate experiments are also provided to verify our claim.

Furthermore, we will carefully re-organize Section 1 of the paper to explain the motivation and contribution better, as well as provide a briefer summary of the core idea. Also see Sec 1.1 of the letter for clearer explanation on the motivation and contribution of CAG-Net.

Q2: Both image de-raining task (raindrop/rainstreak removal) and image de-noising task can be interpreted as image-to-image translation task (e.g., rainy/noisy-to-clean image translation), thus can be solved by any image-to-image translation network (e.g., pix2pix). CAG-Net is also a general image-to-image translation network, thus can be simply trained with different datasets to solve different image-to-image translation problems. See Sec1.1 of the letter to find out the tasks that have been successfully tacked by our CAG-Net.

Q3: The log calculation is used to balance the two different loss terms in Eq (9), and the corresponding reference on explaining such design will be provided in the camera-ready version.

Q4: In Eq (12), α_1 and α_3 are determined by the contributions of different pyramidal levels. For α_2 and α_4 , they are chosen from (0.0001, 0.001, 0.01, and 0.1), and we found that with enough training epochs, the PSNR deviation is only ± 0.002 on RD-Data and RS-Data, and we set $\alpha_2 = \alpha_4 = 0.01$ because this set results in the fastest convergence speed.

Q5: Due to the space limitation, we have to reduce the size of the visual results. For fair comparison, more examples with very large sizes are provided in the supplementary material. Besides, more examples will be released online after paper acceptance.

Q6: Such phenomenon can be interpreted as very light over-deraining, and this will be analyzed by adding a new Section (Limitation Study) to the paper in the camera-ready version. Besides, as can be seen from more examples in Fig. R4, this is absolutely not a general phenomenon.

Fig. R4. Visual comparison of raindrop removal task.

References

[1] Qian, R., Tan, R.T., Yang, W., Su, J. and Liu, J., 2018. 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).

[2] Yasarla, R. and Patel, V.M., 2019. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8405-8414).

[3] Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q. and Lau, R.W., 2019. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 12270-12279).

[4] Li, R., Cheong, L.F. and Tan, R.T., 2019. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1633-1642).

[5] Hu, X., Fu, C.W., Zhu, L. and Heng, P.A., 2019. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8022-8031).

[6] Li, Y., Liu, S., Yang, J. and Yang, M.H., 2017. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3911-3919).