1.1 Motivation and Contribution:

Existing de-raining networks usually suffer from (1) being overfitted to only some specific rain types covered by the training sets and (2) being unable to perform well on real rainy images. The reason can be interpreted as: the representation learned by the encoder in the latent space is a mixture of rainy related factors (RRF) and background related factors (BRF), with the RRF playing a dominate role (see Fig. R1) especially for images with heavy rains, thus resulting in the valuable background factors cannot be well used by the decoder to reconstruct the clean images. To resolve this and to learn an explainable, controllable, and generalizable de-raining model, the idea of isolating the RRF and BRF in the latent space such that only the BRF is used for reconstructing the rain-free image is for the first time proposed and implemented by a representation disentanglement design. For the network design, a novel weakly-supervised multi-task learning framework is designed involving four simple subtasks on clean-to-clean translation, rainy-to-clean translation, clean-to-rainy translation, and rainy-to-rainy translation (see Fig. 3 of the paper), and an elegant knowledge transfer strategy driven by a novel regularized discriminator design is formulated to enable desired disentangled representation learning. After training, the model has been demonstrated to be able to (1) deal with different rainy conditions, (2) generalize well to unseen examples, and (3) most importantly perform well on real rainy images. What's more, the model can be easily adapted to solve other problems such as (1) using the proposed adversarial loss formulation to improve the performance of many GAN models, (2) using the proposed framework directly to learning disentangled representation in a weakly-supervised manner on other datasets, and (3) the clean-to-rainy translation branch can be used for rainy style transfer, which can be applied on real rainy images for creating rainy and clean image pairs to enable the training of a de-raining model on real-world images in a supervised manner.

1.2 How DRLE-Net works:

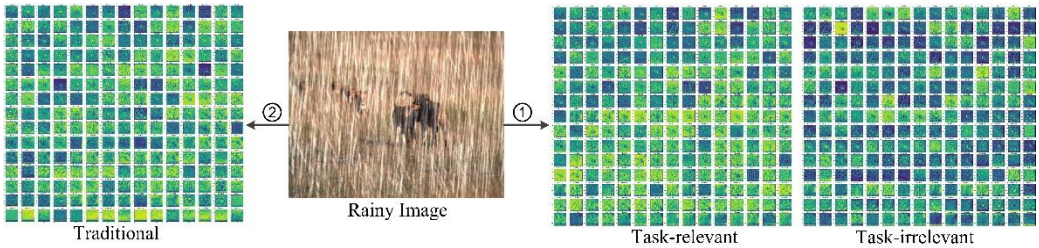

Fig. R1 Visualization of activations (before ReLU) for the bottleneck layer. 1 denotes that the feature maps are from the proposed DRLE-Net, and 2 denotes that the feature maps are from a traditional encoder-decoder (the baseline model described in Table 1) de-raining network.

As can be seen from Fig. R1, the task -relevant/-irrelevant factors can be clearly separated by the proposed framework, thus only the task-relevant factor will be used to reconstruct high-quality de-rained image. In contrast, the traditional network design will learn a mixed latent factor with most features related the task-irrelevant factor, inevitably resulting in under-deraining results. Besides, for rainy image A and B, with our DRLE-Net the task-relevant factor of A and task-irrelevant factor of B can be combined, enabling the rainy

style of B transferred to A (examples will be shown in the GitHub page after paper acceptance).

1.3 How the improved discriminator works:

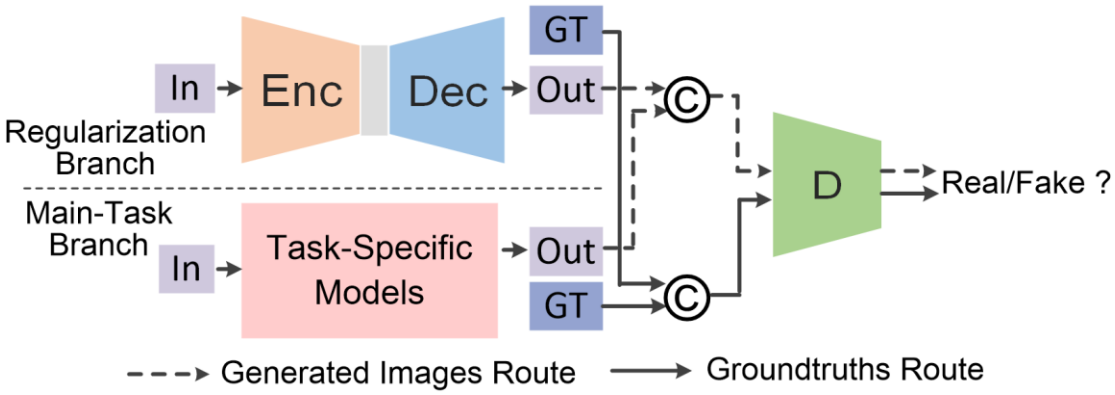

Fig.R2: Illustration of a general generative model design with the improved adversarial loss setup.

Take Fig.R2 of the paper as example, by only combining the "main-task" branch (MTB) with discriminator (D), it is a traditional GAN and the D gets too successful that the generator gradient vanishes and learns nothing. However, by adding the regularization branch (RB), which shares the same task as MTB but take the groundtruth as input, the output from RB will be much more like the groundtruth. By concatenating the output from MTB and RB together as input, the D will not easily recognize the combined input as fake, thus can be trained in a more stable and easier manner, resulting in better image generation by the MTB. Such mechanism is also helpful for the proposed CERLD-Net because it can help guarantee a more stable knowledge transfer procedure, thus resulting in better factor disentanglement. Despite the simplicity of the novel discriminator design, elaborate experiments have been carried out (Sec4.3 and 4.5 of the paper) to demonstrate its effectiveness, and we will add the above analysis to the paper in a future version for better motivation understanding.

2.1 Response to specific comments (#1):

Q1: The FUS layer is similar to a Squeeze-and-Excitation block, and GAP is conducted in the spatial dimension. The FUS layer is designed for feature fusion because it is specifically formulated to make the features (connected by the skip-connection) to be more compatible, as done in ERL-Net.

Q2: The groundtruth (GT) distribution map is generated by defining the mean pixel value as threshold, and this thresholding strategy is too coarse to generate inaccurate GT. Taking the inaccurate GT as input, the errors will also be propagated to the de-raining network thus producing unsatisfactory results. On the other hand, for our design by using a DMG-Net to generate distribution map, the DMG-Net will receive two supervisions including (1) the L1 distance between the output and the inaccurate GT and (2) the indirect de-raining loss. By such way, the negative loss induced from the inaccurate GT will be mitigated by the de-raining loss in an implicit manner, thus producing better de-raining results. Driven by this,

we also set the weight for the L1 loss to be as small as 0.05 to avoid the changing of gradient of DMG-Net being dominated by the inaccurate GT.

Besides, following the setup of the ablation study in Sec4.3, we also compare the results by concatenating the distorted images with (1) GT or (2) the output from DMG-Net. Results are shown in Table R1.

Table R1: Results by using different information for task-relevant factor enhancement.

Setup	RS-Data	RD-Data
DMG-Net	30.92/0.935	32.01/0.938
GT	30.15/0.928	31.56/0.932

Results in Table R1 demonstrates the superiority of our design over the setup of directly using GT distribution map. We will add the results and analysis to Sec 4.3 of the paper in a future version.

2.2 Response to specific comments (#2):

Q1: Both the formulation and the network structure are novel and different from existing models. Please see Sec 1 of this letter and the comments of R1 on the novelty and contribution of our paper.

Q2: Our DRLE-Net differs from Deblur-Net [1] from the following two aspects: (1) we propose to achieve the factor disentanglement using paired data under the weakly-supervised setup and the Deblur-Net achieve this under unsupervised setup. Taking de-raining task as example, for paired data setup, the difference only reflects in rainy components while the background is the same, thus guarantee a very precise factor separation with a well-defined framework (e.g., our DRLE-Net). Differently, Deblur-Net uses unpaired data for disentanglement via decomposing the image into style and content space. However, this is problematic because the unpaired images differ not only in style but also in content, thus resulting in a very confusing separation. By using these two frameworks for rainy images component disentanglement, we visualize the feature map of content space from Deblur-Net and the background factor from our DRLE-Net. As shown in Fig. R2, our model achieves a much better disentanglement than Deblur-Net, the content code of which are very confusing.

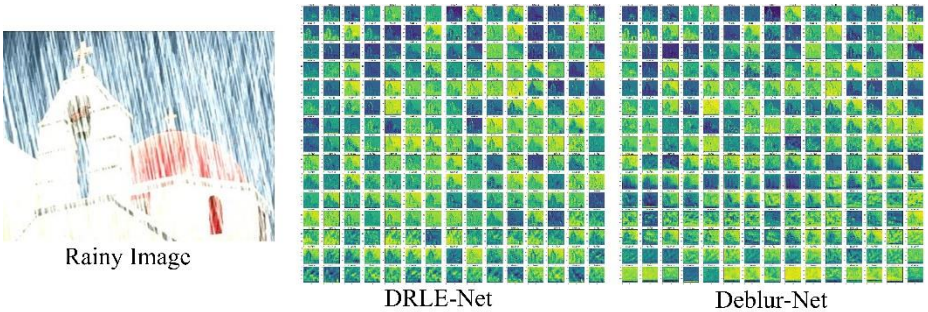

Fig. R2 Visualization of activations (before ReLU) for features from the output layer of task-relevant factor encoder in DRLE-Net and features from the output layer of content encoder in Deblur-Net.

(2) Apart from the theoretical analysis, we also use the Deblur-Net to do unsupervised

image de-raining task on RD-Data and RS-Data by using random shuffling to break the pair relationship of existing dataset. The comparative results are as shown in Table R2.

Table R2: Results by using different disentanglement models.

Models	RS-Data	RD-Data
Deblur-Net	26.84/0.896	28.72/0.901
DRLE-Net	30.92/0.935	32.01/0.938

As can be seen from Table R2 and as expected, the unsupervised Deblur-Net is outperformed by our weakly-supervised DRLE-Net by a large margin, confirming our analysis on the deficiency of Deblur-Net for factor disentanglement. We will add the results and analysis to the paper in a future version.

Q3: For the ablation study results in Table 1 of the paper, one may question that the performance improvement may come from more parameters brought by adding different modules into a baseline model. To answer this, we try to minimize the parameters by replacing the traditional convolution used in different modules (e.g., FUS, FDM, IDM, and DMG-Net) as separable convolution and replacing the FC layer as a separable convolution layer followed by a GAP layer, and we found that the average PSNR drops are within 0.1dB, and our DRLE-Net still achieves new SOTA results when compared with other competitive models on both raindrop and rainstreak removal tasks. Detailed results and the corresponding codes will be provided after paper acceptance.

Q4: For the missing references, we will (1) add them to “Related Work” and (2) add both the quantitative and qualitative results by these references to Sec 4.4 of the paper.

2.3 Response to specific comments (#3):

Q1: Both the formulation and the network structure are completely new and different from existing models. Please see Sec 1 of this letter and the comments of R1 on the novelty and contributions of different components in our de-raining model.

Q2: (1) For the residual-aware models, they are using the concept of disentanglement in the image domain by simply defining the rainy image as the linear superposition of rain layer and background layer, and de-raining is achieved by formulating a network to learn the rain layer which is then used together with the rainy image to obtain the residual image. However, the simple linear superposition assumption is too coarse to describe the formation of images with diversified rainy components (i.e., the real-rainy images), thus leading the trained residual-aware models to be (i) easily overfitted to only some specific rain types covered by the training set and (ii) unable to perform well on real rainy images. In contrast, the factor disentanglement in DRLE-Net is achieved in the latent space with high semantics, and two networks are defined to learn the distribution of each factor in a non-linear manner. Consequently, as demonstrated by the results in Sec 4.4 of the paper, our DRLE-Net is able to perform well on both synthetic and real images with diversified rainy conditions, and outperform the residual-aware models by a large margin; (2) For raindrop segmentation aware models, they are not related to the disentanglement design in either the image domain or the semantic feature domain. Alternatively, they are formulated by using the guidance information (e.g., rainy components segmentation map in AGAN/ JORDER and density label in DID-MDN) to drive the filter in the de-raining network to be more activated on the

rainy regions, such that all rainstreaks/raindrops can be attended for producing better de-raining results without artefacts of under-deraining. Good results of such models come from the usage of the guidance information to learn better semantic feature that will be used by the decoder to reconstruct the clean images. For a de-raining task, only the background related features are useful for reconstructing the rain-free images. However, without a factor separation process, the semantic features cover factors related to both rainy component (task-irrelevant factor) and background component (task-relevant factor). Consequently, the guidance information will take effect on both the task -relevant and -irrelevant factor learning procedure. To avoid this and find a better solution for only improving the learning of the task-relevant factor, we formulate a novel framework for joint factor disentanglement and enhancement. Results in Sec 4.4 of the paper also demonstrate the superiority of our framework over the segmentation aware models (JORDER/URML and AGAN). (3) For the image-decomposition based models (DSC and GMM in Table3), the rainy image is separated into two subspaces represented by Gaussian Components or dictionaries, then optimization methods (e.g., sparse coding or EM algorithm) are used to derive the representation of rainy and background components. This procedure can also be interpreted as disentanglement, and DRLE-Net is in spirit similar to such process by using a neural network for separation. However, compared to the image-decomposition based models, DRLE-Net shows many advantages such as (a) being easier to be optimized and much faster during inference, (b) producing much better de-raining results, and (c) showing high potential of being used for solving other problems as analyzed in Sec 1.1 of the letter.

Reference:

[1] Lu, B., Chen, J.C. and Chellappa, R., 2019. Unsupervised domain-specific deblurring via disentangled representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10225-10234).