RAIN-PRIOR INJECTED KNOWLEDGE DISTILLATION FOR SINGLE IMAGE DERAINING

Yuzhang Hu, Wenhan Yang, Jiaying Liu , Zongming Guo*

Wangxuan Institute of Computer Technology, Peking University, Beijing, China

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897845

ABSTRACT

This paper makes efforts in improving the efficiency of deep networks for single image deraining with a newly proposed knowledge distillation framework. Specifically, we propose a rain-prior injected distillation scheme to transfer the knowledge from a large-scale teacher network to a more compact student network. Previous works directly calculate the distillation loss between the features extracted from the student and teacher networks. Differently, our distillation scheme adaptively removes the noisy background patterns by calculating the distillation loss based on the residual feature, which is inferred from the features extracted from the rain and ground truth images. This residual operation makes the student network focus on transferring only the knowledge on the rain streaks instead of the background, which facilitates more effective distillation results. Furthermore, our method can be applied to reduce both the network size and the deraining recurrence stage, which makes it a plug-and-play module that can be integrated into diverse existing deraining methods. Experimental results prove the efficiency of our method to build an efficient deraining network and the superiority over existing distillation methods.

Index Terms— Single Image Deraining, Knowledge Distillation, Rain-Prior Injection, Recurrent Distillation

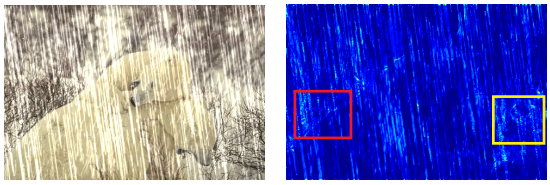
1. INTRODUCTION

Images captured on rainy days suffer from visual degradation, which leads to low visual quality and failures in computer vision tasks. With the development of deep learning, there are a series of methods [1, 2, 3, 4] for the goal of rain removal with deep neural networks. To achieve the goal of better deraining performance, there are two trends in the development of deraining methods. First, the number of the parameters of the network is becoming larger and larger, and the architecture of the network becomes more complex. In this way, the capacity of the network is improved to deal with more diverse rain distributions and provide better performance. Second, to deal with the heavy rain, multiple recurrences are adopted in [5, 6]. In each recurrence, the input rainy image is processed with a deraining network, and the deraining result is viewed as the input to the next recurrence. This framework is better at dealing with heavy rain. Although significant improvement has been achieved, these two trends lead to larger storage space and more computational costs, which are unfriendly to real applications. It is of great significance to build a more efficient deraining model for practical applications.

Knowledge distillation provides a feasible direction to address the issue. This technique is first proposed for image classification [7], to transfer the knowledge from a more large and powerful deep network (*i.e.*, teacher network) to a smaller network (*i.e.*, student network). It is believed that the output or the intermediate feature of the teacher network contains rich supervision information out of the ground truth, which can play the role of guidance to assist the student network to achieve better performance. There are also some explorations of the knowledge distillation in low-level visions. In [8], four different distillation losses are imposed on the intermediate feature to distill an image super-resolution network and the efficiency of these different distillation losses is compared. In [9], a more complex distillation loss is designed by introducing the affinity matrix, which reflects the spatial correlation of the intermediate feature. However, these methods only concentrate on designing different loss functions, which ignore the physical property of the degradation and might not be effective for single image deraining. As shown in Fig. 1 (b), except for the rain information, there are also remaining noisy background patterns in the intermediate features obtained by the network. As a result, if the distillation loss is directly imposed on the intermediate features, this mixture makes it difficult for the student network to learn useful guidance and leads to poor distillation results.

In this paper, we propose a rain-prior injected distillation scheme to address the issue. Different from existing distillation approaches in low-level vision, we explore a more effective distillation mechanism instead of designing a more complex distillation loss. Under our scheme, both the rainy image and the corresponding clean image are fed into the network. The intermediate feature of the rainy image will be subtracted by the intermediate feature of the clean image to obtain the residual feature. In this way, the noisy background patterns in the intermediate feature of the rainy image can be suppressed and only rain streak information remains as shown

*Corresponding Author. This work was supported by the National Natural Science Foundation of China under Contract No.62172020 a research achievement of Key Laboratory of Science, Techonology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).



(a) Rainy Input (b) Feature of Rainy Image

(c) Feature of Ground Truth (d) Residual Feature

Fig. 1: Visualization of the intermediate feature of different inputs. Clean background and rain streak patterns are mixed in the feature of the rainy image. The residual feature, on the contrary, contains only rain streak information with few background patterns.

in Fig. 1 (d). The distillation loss is finally imposed on the residual features of the student and teacher networks. Thus, a more efficient student network can learn rain-related knowledge from a stronger teacher network. Besides, we further extend our scheme to adapt to the distillation from a recurrent deraining network to a single-stage deraining network, which makes our scheme plug-and-play for diverse deraining methods. With the well-designed rain-prior injected distillation scheme, we obtain a 0.95dB improvement in terms of Peak Signal-to-Noise Ratio (PSNR) for more efficient rain removal.

2. RAIN-PRIOR INJECTED KNOWLEDGE DISTILLATION

In this section, we introduce the proposed rain-prior injected knowledge distillation scheme in two parts. First, we introduce the basic framework of the distillation process between vanilla deraining networks. Then, we extend our framework to distill the knowledge from a recurrent teacher network to a single-stage student network.

2.1. Basic Distillation Framework

Fig. 2 shows the pipeline of our rain-prior injected knowledge distillation scheme. The Teacher Net and the Student Net are pretrained so both of them can perform image deraining. Then, we divide a deraining network into two parts. The Feature Extraction part maps the input image into the intermediate feature space, and the Decoder part maps the intermediate feature back to the image space to obtain the deraining results. While for the Teacher Net, network structure is more complex with more parameters, so it achieves better deraining performance than the Student Net. Our goal is to improve the

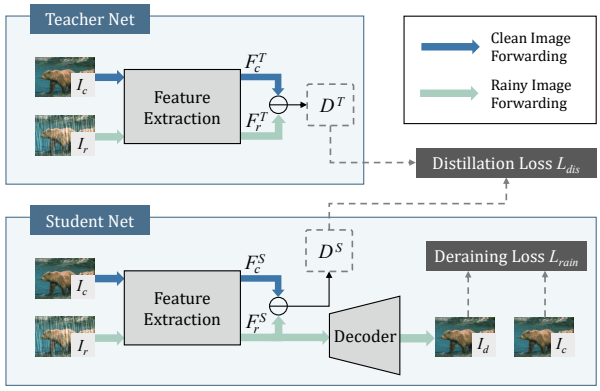


Fig. 2: The pipeline of the rain-prior injected knowledge distillation. A lightweight student network is distilled from a more powerful teacher network. Both rainy images and clean images are fed into two deraining models to obtain the residual features, based on which the distillation loss is further constructed to guide the student network to learn useful knowledge from the more powerful teacher network.

performance of the Student Net through the distillation process with the guidance of the intermediate feature from the Teacher Net.

Our distillation scheme adaptively removes the noisy background patterns. The distillation loss is inferred based on the residual feature inferred from the features extracted from the rainy and ground truth images. More specifically, first, as the Clean Image Forwarding branch in Fig. 2 shows, a clean image I_c is fed into both the Teacher Net and the Student Net to obtain the intermediate feature F_c^T and F_c^S, respectively. F_c^T and F_c^S only contain background information because there is no rain in the input I_c. After that, the corresponding rainy image I_r is fed into both the Teacher Net and the Student Net to obtain the intermediate feature F_r^T and F_r^S from the Rainy Image Forwarding branch as shown in Fig. 2. F_r^T and F_r^S contain the information of both background patterns and rain streaks. What is more, F_r^T includes richer information compared with F_r^S due to the powerful capacity of the Teacher Net. Then, the intermediate feature of the rainy image is subtracted with that of the clean image to obtain the residual feature D^T and D^S. As shown in Fig. 1 (d), the residual feature mainly contains rain streak information with fewer background patterns. Then, following [8], all the residual feature is aggregated to a one-channel feature as follows:

$$\widetilde{D} = \left(\sum_{c=1}^{C} D_{(c)} \right)^2, \tag{1}$$

where C is the channel number and $D_{(c)}$ is the c-th channel of the residual feature D. The distillation loss is imposed on the aggregated residual feature as follows:

$$L_{dis} = \left| \widetilde{D}^T - \widetilde{D}^S \right|. \tag{2}$$

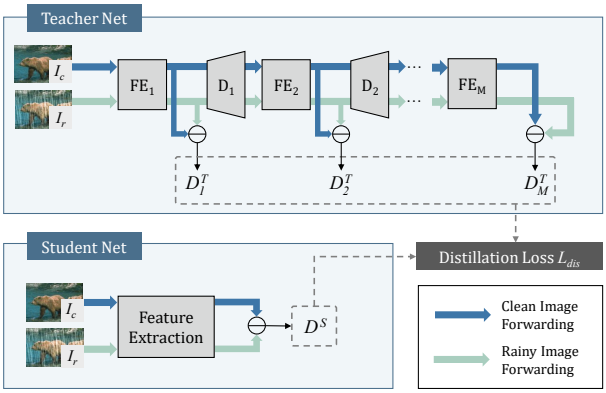


Fig. 3: The distillation pipeline for recurrent Teacher Net. The residual feature of each recurrence is utilized to transfer the knowledge to the Student Net.

Besides, the intermediate feature F_r^S is fed into the subsequent Decoder to obtain the deraining result I_d of the Student Net, and the deraining loss is imposed with the corresponding clean image as follows:

$$L_{rain} = SSIM(I_d, I_c), \tag{3}$$

where SSIM stands for the Structural Similarity Index Measure [10] metric. The Student Net is trained with the total loss as follows:

$$L_{total} = L_{rain} + \lambda L_{dis}, \tag{4}$$

where λ is the weight to balance two terms, which is set to 0.1 in our implementation. The parameters of the Teacher Net are fixed during the distillation process.

2.2. Distillation for Recurrent Teacher Net

We further explore extending the above-mentioned distillation framework to distill the knowledge from a recurrent Teacher Net to a single-stage Student Net. As shown in Fig. 3, the Teacher Net consists of multiple recurrences. FE$_i$ stands for the Feature Extraction part of the i-th recurrence and D$_i$ is the Decoder to generate the deraining result of the i-th recurrence. This deraining result is viewed as the input of the next deraining recurrence. M is the total number of all recurrences. Under this structure, the residual feature of each recurrence is calculated in the way described above and the distillation loss in Eqn. (2) is updated as follows:

$$L_{dis} = \sum_{i=1}^{M} \left| \widetilde{D}_i^T - \widetilde{D}^S \right|, \tag{5}$$

where $\widetilde{D}_i^T$ is the one-channel aggregated residual feature of the i-th recurrence. The total loss to train the Student Net is the same as Eqn. (4) and the deraining loss imposed on the Student Net is not shown in Fig. 3 for simplicity.

3. EXPERIMENTAL RESULTS

3.1. Experiment Setting

We evaluate the efficiency of our methods in two parts. In the first part, we perform the distillation pipeline in Sec. 2.1 between networks without the recurrent structure, namely single-stage distillation. In the second part, we apply the distillation pipeline in Sec. 2.2 to distill the knowledge from a recurrent Teacher Net to a single-stage Student Net, namely recurrent distillation. Two commonly used datasets, Rain100H [2] and Rain1200 [3] are adopted in the evaluation.

Network Configuration. We combine the recursive residual group [11] as the Feature Extraction module and a single convolutional layer as the Decoder module to build a deraining network. A series of deraining networks are obtained by setting the channel number of the intermediate feature and the deraining recurrence as different values. We build one Teacher Net and two Student Net for each part of the evaluation described above. For simplicity, the Teacher Net is denoted as TN and two Student Net is denoted as SN1 and SN2. The detailed network configurations of two evaluation parts are shown in Tab. 1 and Tab. 2, respectively.

Training Details. The network is implemented in PyTorch and AdaMax [12] is used as the optimizer. We crop 64×64 patches from each image during the training stage with the batch size set to 8. The learning rate is first set to 10^{-4} and drops automatically to 10^{-6} with a decay rate of 0.5. We train the Student Net for 100 epochs on an RTX 2080 GPU.

Table 1: Configurations of the networks in single-stage distillation.

Network	#channel	#parameter
TN	64	10.6M
SN1	16	333K
SN2	8	167K

Table 2: Configurations of the networks in recurrent distillation.

Network	#recurrence	#channel	#parameter
TN	3	64	31.8M
SN1	1	64	10.6M
SN2	1	32	2.65M

3.2. Overall Performance

Tab. 3 and Tab. 4 show the performance improvement of the single-stage distillation and the recurrent distillation, respectively. It can be observed that both Student Nets of different sizes can benefit from our distillation scheme. Specifically, up to 0.95dB improvement can be obtained for the Student Net with the distillation of a recurrent Teacher Net.

Table 3: Distillation results with single-stage Teacher Net.

Dataset	Metrics	TN	SN1		SN2	
			w/o distillation	w distillation	w/o distillation	w distillation
Rain100H	PSNR	30.11	27.57	28.04	26.84	27.42
	SSIM	0.9169	0.8814	0.8872	0.8681	0.8739
Rain1200	PSNR	32.84	31.71	32.00	31.12	31.58
	SSIM	0.9229	0.9095	0.9119	0.9047	0.9084

Table 4: Distillation results with recurrent Teacher Net.

Dataset	Metrics	TN	SN1		SN2	
			w/o distillation	w distillation	w/o distillation	w distillation
Rain100H	PSNR	33.03	30.11	30.87	28.99	29.94
	SSIM	0.9466	0.9169	0.9252	0.9010	0.9127
Rain1200	PSNR	33.27	32.55	32.88	32.25	32.56
	SSIM	0.9295	0.9213	0.9232	0.9173	0.9202

Table 5: Distillation results of existing deraining methods.

Metrics	PredNet		RESCAN	
	baseline	distilled	baseline	distilled
PSNR	23.14	24.07	23.10	23.44
SSIM	0.8402	0.8441	0.8180	0.8196

We further show the visual improvement of the deraining result with the distillation. As shown in Fig. 4, there are remaining rain streaks and blurry edges in the output of a Student Net before the distillation. In comparison, the deraining result is of better visual quality after the distillation.

3.3. Verification on Existing Deraining Methods

In order to prove the generalization capacity of our method, we take two existing deraining methods, PredNet [5] and RESCAN [6], as the Student Net and distill them with the Teacher Net in Tab. 1. As shown in Tab. 6, our distillation framework can be applied for various deraining methods with moderate performance improvement.

3.4. Comparison with Existing Distillation Methods

We choose three existing distillation methods to compare with our rain-prior injected distillation scheme. FitNet [13] proposes to directly align the intermediate feature of the Teacher Net and the Student Net. SRKD [8] utilizes the statistical map of the intermediate feature for distillation. FAKD [9] calculates the spatial similarity of the intermediate feature before the distillation process. The comparison results are shown in Tab. 6. Our method obtains the biggest performance improvement compared with other methods.

(a) Before Distillation (b) After Distillation

Fig. 4: Visual results of the distillation. The deraining result of the network after the distillation is of better visual quality.

Table 6: Comparison with existing distillation methods.

Method	baseline	FitNet	SRKD	FAKD	Ours
PSNR	26.81	26.96	27.21	27.33	**27.42**
SSIM	0.8681	0.8695	0.8722	0.8721	**0.8739**

4. CONCLUSION

In this paper, we propose a rain-prior injected knowledge distillation framework to improve the deraining efficiency of deep networks. The removal of the background patterns makes the student network focus more on rain streak information instead of the background information for an efficient knowledge transfer. The framework is also extended from a single-stage model to a recurrent model to achieve better generalization capacity. Experimental results prove the efficiency of our framework and its superiority to other distillation methods.

5. REFERENCES

[1] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley, "Removing rain from single images via a deep detail network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3855–3863.

[2] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan, "Deep joint rain detection and removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1357–1366.

[3] He Zhang and Vishal M Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 695–704.

[4] Wenhan Yang, Jiaying Liu, Shuai Yang, and Zongming Guo, "Scale-free single image deraining via visibility-enhanced recurrent wavelet learning," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2948–2961, 2019.

[5] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3937–3946.

[6] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 254–269.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong, "Image super-resolution using knowledge distillation," in *Proceedings of the Asian Conference on Computer Vision*, 2018, pp. 527–541.

[9] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia, "FAKD: Feature-affinity based knowledge distillation for efficient image super-resolution," in *Proceedings of the IEEE International Conference on Image Processing*, 2020, pp. 518–522.

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[11] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, "Learning enriched features for real image restoration and enhancement," in *Proceedings of the European Conference on Computer Vision*, 2020.

[12] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2014.

[13] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.