From Sky to the Ground: A Large-scale Benchmark and Simple Baseline Towards Real Rain Removal

Supplementary Material

1. Overview

In this supplementary material, we present more analysis and experimental results.

- We introduce the phenomenon of ground splashing and its negative effects. (Section 2)
- We conduct analysis among existing paired real rain benchmarks. (Section 3)
- We provide additional ablation study of proposed RLRTR and SCD-Former. (Section 4)
- We show more experimental results of datasets generalization and high-level tasks. (Section 5)
- We discuss the time consuming, user study and the detailed solution of RLRTR. (Section 6)

2. What is Ground Splashing Rain?

Ground splashing rain is commonly observed under heavy rain in the real world. It is caused by the reflection of rain hitting on the ground and presents as dense point-shape texture, splashing droplets or water waves, as shown in Fig. 1, which interferes the visibility of the traffic signs on the ground, lane lines, speed limits, crossing road, etc. The splashing rain on the ground causes tremendous negative effects on downstream vision tasks such as objects detection and segmentation.

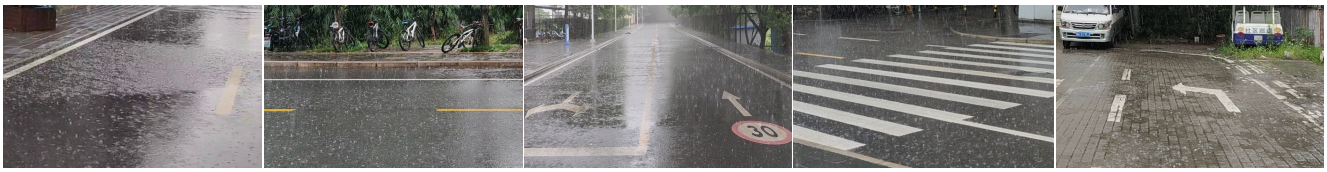


Figure 1. Typical examples of ground splashing rain. It presents as dense and messy point-shape texture, splashing droplets or water waves on the surface of road, which ruins the visibility of traffic signs and brings negative interference to vision systems such as self driving.

3. Benchmark Analysis

In this section, we make a brief comparisons of existing paired real rain benchmarks SPA-data, GT-Rain and RealRain-1K and further analyse their limitations in Fig. 2. Then we explain how to overcome the bottleneck of real rain removal by our proposed benchmark, LHP-Rain.

3.1. SPA-data

SPA-data is the first paired real rain dataset which contains 170 sequences collected from Internet and real world. The feature of SPA-data is the strictly paired rain/clean images which enhances the learning of real rain. The dataset releases the image patches (256*256) for training. Typical examples of SPA-data are illustrated in the first row of Fig. 2. On the one hand, SPA-data contains limited rain categories and narrow views. As shown in Fig. 2(a), most rainy images from SPA-data have similar rain streaks and narrow views which focus on the roof, eave, building and plants without clear objects. On the other hand, sequences collected from Internet contain obvious watermarks, which disturb the original rain and image appearance. Therefore, SPA-data has limited generalization on other real rain datasets.

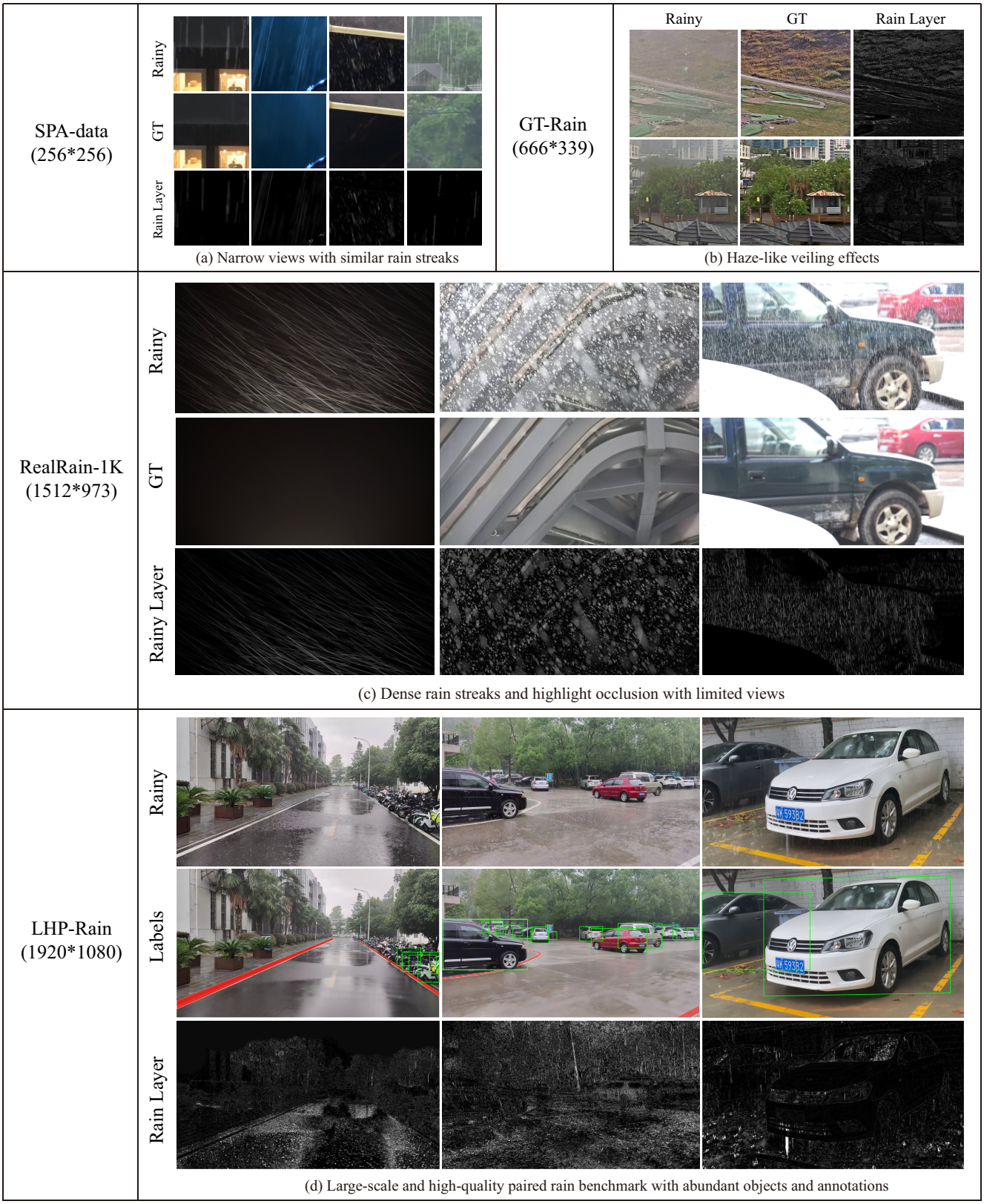


Figure 2. Benchmark analysis of existing paired real rain datasets. We choose the typical examples from existing four benchmarks: SPA-data, GT-Rain, RealRain-1K and LHP-Rain. The Rainy, GT and Rainy layer images are illustrated.

Figure 3. Ablation study of the robust low-rank tensor recovery model. From left to right, the first column is the original rainy frame and the remaining three columns represent the model without non-local self-similarity prior, the model without local smoothness prior and the model with both.

3.2. GT-Rain

GT-Rain is a paired real rain dataset proposed recently, which contains 202 sequences collected from Youtube live stream with average resolution 666*339. GT-Rain leverages the static surveillance webcams to produce rain/clean image pairs. As shown in Fig. 2(b), it contains rain streak and haze-like veiling effect due to the long interval between rainy and clean images, which effectively enriches the abundance of rain categories. However, the poor image quality issue limits the performance of GT-Rain. Due to the compression of online videos, the poor image quality leads to the change of original rain appearance, which is suppressed or even unseen in the rainy images.

3.3. RealRain-1K

RealRain-1K is a very recently proposed datasets as well, which collects 1120 sequences from Internet and real world. To solve the problem of low resolution, the average resolution of RealRain-1K increases to 1512*973, which is larger than SPA-data and GT-Rain. Typical examples are illustrated in Fig. 2(c). The strength of RealRain-1K is that it contains clear rainy images including abundant rain streak, veiling effect and highlight occlusion. However, most rainy images of the dataset focus on the dense rain streaks in front of the camera while ignoring the details of background. It is unreasonable because the rain appearance is entangled with environment [8]. Therefore, the background of RealRain-1K suffers from out-of-focus blur and even contains nothing without clear objects, which limits the further applications on high-level tasks.

3.4. LHP-Rain

To solve the limitations of existing datasets, we proposed a large-scale and high-quality paired real rain benchmark, LHP-Rain. Typical examples are illustrated in Fig. 2(d). First, the LHP-Rain contains **diverse rain categories with very large-scale**, including 3000 video sequences with over 1 million frame pairs. Apart from the conventional streak, veiling and highlight occlusion, our benchmark contains representative the challenging case **ground splashing rain**. Moreover, the LHP-Rain is collected by the smartphone with high-resolution (1920*1080) and abundant objects for self-driving and surveillance. Five typical objects including person, car, bicycle, motorcycle and bus are annotated by bounding box with 326,961 instances totally. For lane segmentation, we annotate 24,464 lane masks to evaluate the effect of rain splashing removal. Last but not least, the **high-quality ground truth** enhances the discriminative representation of rain feature learning.

4. Ablation Study

4.1. Ablation Study on Video Deraining: RLRTR

Effectiveness of Non-Local Self-similarity. The non-local prior is used to explore the self-similarity of the spatial dimension. To verify its effectiveness, we perform an ablation study on Fig. 3. It can be observed that the results obviously exist the rain residual without the non-local self-similarity prior, implying that relying only on the temporal information is not enough and the non-local low-rank property can boost the deraining effect.

Effectiveness of Total Variation Regularization. The total variation regularization is exploited to characterize the local smoothness of the temporal dimension. As shown in Fig. 3, the results are slightly blurry without the local smoothness prior,

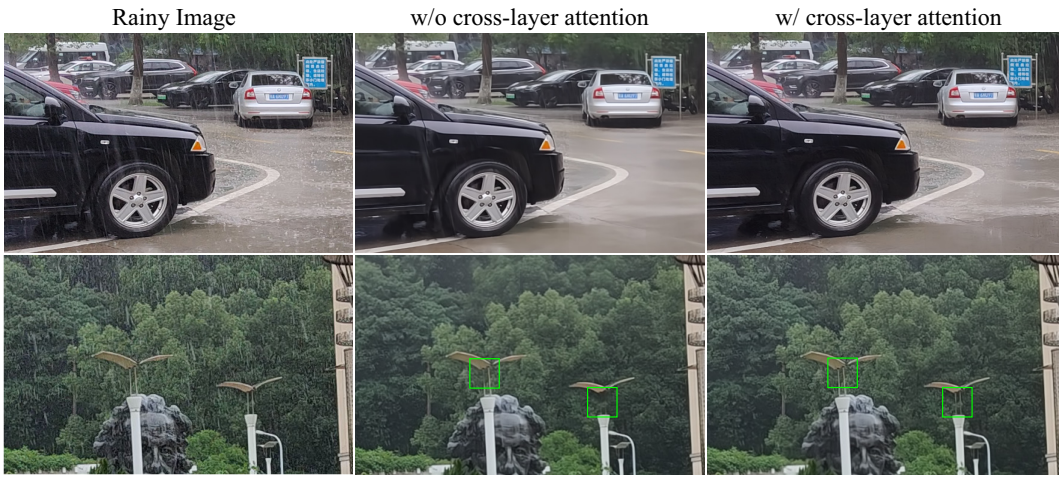


Figure 4. Effectiveness of Cross-Layer Attention. CLA could further remove the challenging rain residual left on the car and prevent the image layer from being destroyed, such as the white holders of street lamps.

Figure 5. Generalization of SCD-Former on different devices. The rainy images (first row) and deraining results (second row) demonstrate that the SCD-Former model trained on LHP-Rain generalizes well on the real rainy images captured by different devices with specific ISP pipelines, iPhone13, Huawei P40, Realme GT2, DSLR, etc.

which demonstrate that it has little effect on the results of restoration and the global low-rank prior of the temporal dimension has been able to capture most of the temporal information.

4.2. Ablation Study on Image Deraining: SCD-Former

Effectiveness of Cross-Layer Attention. In order to further evaluate the effectiveness of cross-layer attention (CLA) module in SCD-Former, more visualization results of ablation study are shown in Fig. 4. Specifically, the residual rain left on the black car could be paid more attention to remove with CLA, because the highly attentive location of rain residual can be obtained by image layer. Moreover, the rain-like structures in the image layer are wrongly removed, such as the white holders on the street lamps. Here CLA could provides an extra prior for enhanced feature representation, which prevents the image layer from being destroyed.

5. Discussion

Generalization on Different Digital Devices. We conduct experiments on the images collected by different devices to evaluate the generalization of SCD-Former. In Fig. 5, SCD-Former is very robust to the data captured by different equipments from smartphones to DSLR, which means that the influence of rain obviously outweighs that of ISP pipelines.

Evaluation of the Diversity of LHP-Rain Benchmark. To evaluate the rain diversity on other real rain datasets, we train

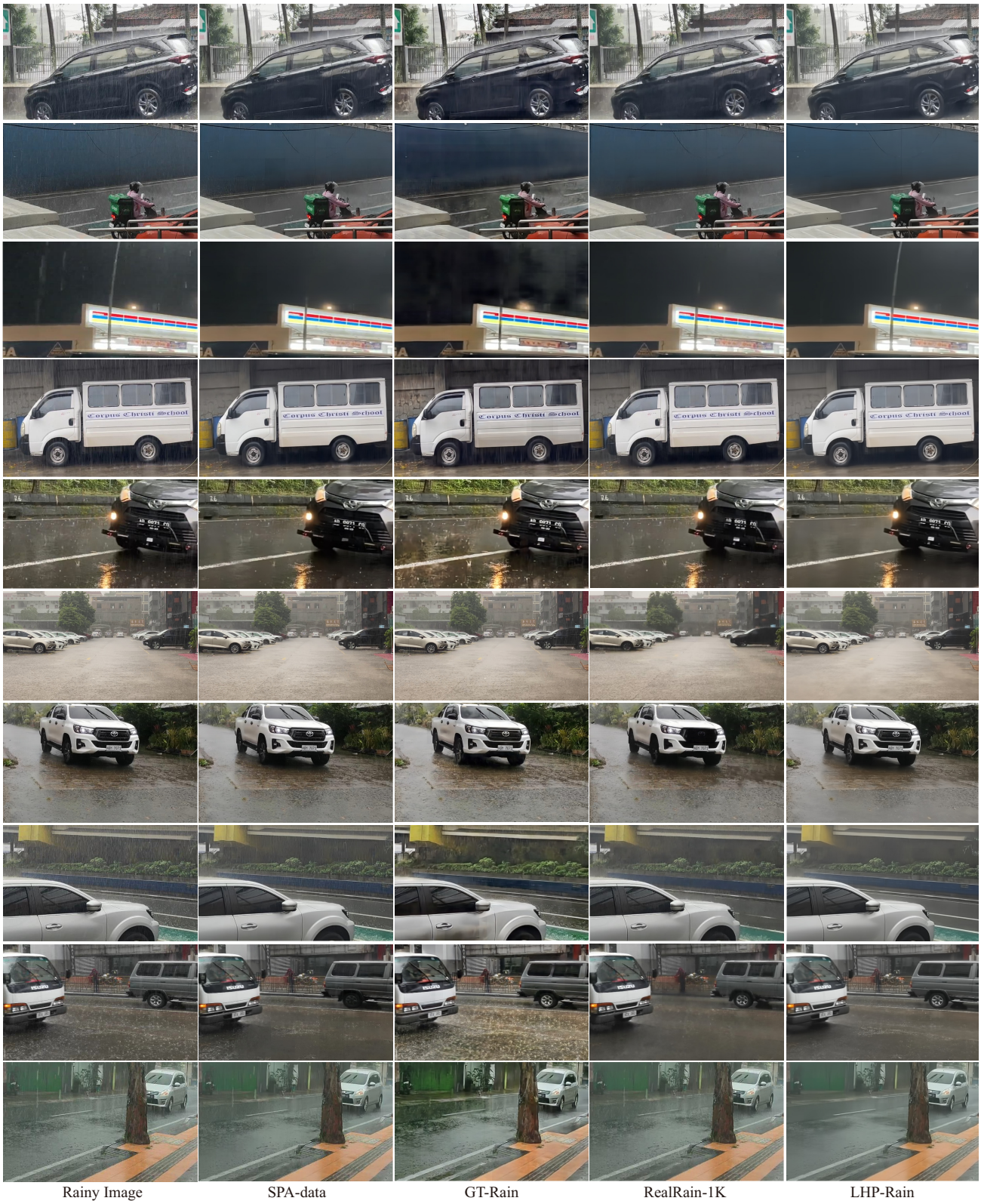
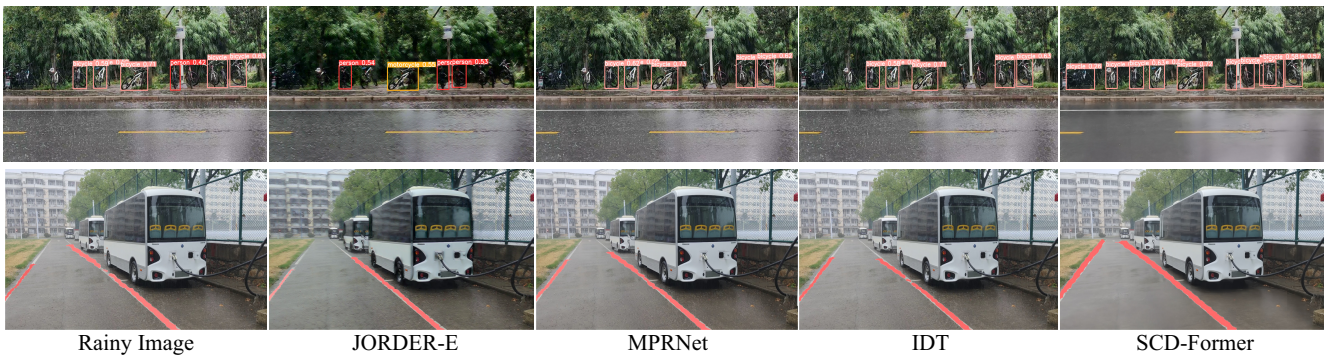


Figure 6. Evaluation of the diversity of the LHP-Rain. We train the same model on different datasets: SPA-data, GT-Rain, RealRain-1K and LHP-Rain, and then test on other datasets. The model trained on LHP-Rain has achieved better deraining results in the real world.



Rainy Image JORDER-E MPRNet IDT SCD-Former

Figure 7. Evaluation of high-level tasks such as object detection and lane segmentation. The result shows that the motorcycle and bikes covered with rain could be further detected after deraining by SCD-Former, and the lane on the ground could be predicted completely due to the removal of ground splashing water.

Table 1. Evaluation of high-level tasks on deraining results.

Method	Det. (mAP)	Gain (Det.)	Seg. (Acc)	Gain (Seg.)
Rainy	0.543	-	0.237	-
GT	**0.633**	**+0.090**	**0.480**	**+0.243**

the same SCD-Former model on different paired real rain datasets: SPA-data, GT-Rain, RealRain-1K and LHP-Rain, and then test on real rain images from Internet. The comparing results are illustrated in Fig. 6. We choose typical and challenging real rain images such as dense rain streaks, veiling effect and ground splashing rain. It is observed that the model trained on SPA-data and RealRain-1K could remove a few rain streaks, but the backgrounds removed as well, such as the wheel of car. In addition, rain residual is still remained in the image layer. Moreover, the model trained on GT-Rain could not generalize on other real rain datasets as well, which enhances the contrast of residual rain in the image layer. The model trained on LHP-Rain has the best deraining performance which simultaneously removes rain streaks, veiling and ground splashing water and preserve the image details. The results strongly support the great generalization of LHP-Rain.

Evaluation on High-Level Tasks. The abundant annotations provided in LHP-Rain could be utilized to evaluate the image deraining results on high-level tasks. We conduct object detection and lane segmentation experiments and the visualization results of high-level tasks are illustrated in Fig. 7. The result shows that the motorcycle and bikes covered with rain could be further detected after deraining by SCD-Former, and the lane on the ground could be predicted completely due to the removal of ground splashing water, which is consistent with the quantitative results reported in the main text. Moreover, we also provide oracle GT in Table 1 with detection mAP/0.633 and lane seg Acc/0.480 while that of rainy is mAP/0.543 and Acc/0.237, respectively. This strongly supports potential of LHP-rain for high-level tasks.

Video Deraining Time Consuming and Results. The proposed RLRTR needs **9.57s on a 256*256*300** video clip with an Intel i7-9700F CPU. Moreover, it is difficult to evaluate the performance of the video deraining results namely the GT quality. We look for 25 volunteers to anonymously rate (1 to 10) the deraining results of 50 scenes, and write a simple scoring software to guarantee justice and equity of rating. The rank-n scores [4] for each method in Figure 8 shows that RLRTR consistently outperforms the competing methods by a large margin.

6. Solution of the Video Deraining RLRTR

The original optimization problem is shown as follow:

$$\begin{aligned}\left\{\hat{\mathcal{B}},\hat{\mathcal{R}},\hat{\mathcal{J}}_i,\hat{\tau},\hat{Q}_i\right\} = \arg\min_{\mathcal{B},\mathcal{R},\mathcal{J}_i,\tau,Q_i} \frac{1}{2}||\mathcal{B}+\mathcal{R}-\mathcal{O}\circ\tau||_F^2+\mu||\mathcal{R}||_1 \\ +\omega\sum_i\left(\frac{1}{\lambda_i^2}||\mathcal{S}_i\mathcal{B}\times_3 Q_i-\mathcal{J}_i||_F^2+||\mathcal{J}_i||_{tnn}\right)+\gamma||\nabla_t\mathcal{B}||_1,\end{aligned} \tag{1}$$

where $\mathcal{O}\in\mathbb{R}^{h\times w\times t}$ is the rainy video, $\mathcal{B}\in\mathbb{R}^{h\times w\times t}$ is the rain-free video, $\mathcal{R}\in\mathbb{R}^{h\times w\times t}$ represents the rains, τ denotes the affine transformation to ensure the rainy video of each frame is pixel-level aligned, $\mathcal{J}_i$ represents the low-rank approximation, $Q_i\in\mathbb{R}^{d\times t}(d\ll t)$ is an orthogonal subspace projection matrix used to capture the temporal low-rank property and $\mathcal{S}_i\mathcal{B}\in\mathbb{R}^{p^2\times k\times t}$ is the constructed 3-D tensor via the non-local clustering of a sub-cubic $u_i\in\mathbb{R}^{p\times p\times t}$ [2].

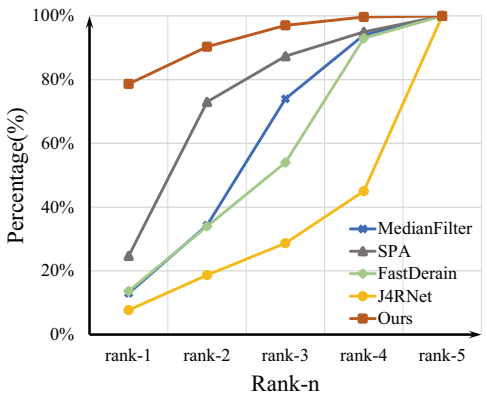


Figure 8. User study of different video deraining methods.

Optimization. To solve this difficult problem, we apply the ADMM [5] to decouple the Eq. (1) into several subproblems.
1) *Affine Transformation* τ: By ignoring variables independent of τ, we can obtain following subproblem:

$$\hat{\tau} = \arg\min_{\tau} \frac{1}{2} \parallel \mathcal{B} + \mathcal{R} - \mathcal{O} \circ \tau \parallel_F^2 . \tag{2}$$

Since $\mathcal{O} \circ \tau$ is a nonlinear geometric transform, it's difficult to directly optimize τ. A common technique is to linearize around the current estimate [6]. After linearizing $\mathcal{O} \circ \tau$, the Eq. (2) can be transform into following formula:

$$\Delta\hat{\tau} = \arg\min_{\tau} \frac{1}{2} \parallel \mathcal{B} + \mathcal{R} - \mathcal{O} \circ \tau - \nabla\mathcal{O}\Delta\tau \parallel_F^2 . \tag{3}$$

Thus, the Eq. (3) can be solved with closed-form solution:

$$\Delta\hat{\tau} = (\nabla\mathcal{O})^{\dagger}(\mathcal{B} + \mathcal{R} - \mathcal{O} \circ \tau), \tag{4}$$

where $(\nabla\mathcal{O})^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $\nabla\mathcal{O}$. And we have:

$$\hat{\tau} = \tau + \Delta\hat{\tau}. \tag{5}$$

2) *Rain Estimation* $\mathcal{R}$: Fixed other variables, we can get following subproblem:

$$\hat{\mathcal{R}} = \arg\min_{\mathcal{S}} \mu \parallel \mathcal{R}\parallel_1 + \frac{1}{2} \parallel \mathcal{B} + \mathcal{R} - \mathcal{O} \circ \tau \parallel_F^2 . \tag{6}$$

The Eq. (6) can be solved by introducing the soft-thresholding operator [5]:

$$\Phi_{\upsilon}(x) = \begin{cases} x - \upsilon, & if \quad x > \upsilon \\ x + \upsilon, & if \quad x < -\upsilon \\ 0, & otherwise \end{cases} \tag{7}$$

Then, we can easily get its closed-form solution:

$$\hat{\mathcal{R}} = \Phi_{\mu}\left(-\mathcal{B} + \mathcal{O} \circ \tau\right). \tag{8}$$

3) *Subspace Projection* Q_i: We enforce the orthogonal constraint on $Q_i^T Q_i = I$ with the following subproblem:

$$\hat{Q}_i = \arg\min_{Q_i^T Q_i = I} \frac{1}{\lambda_i^2} ||\mathcal{S}_i\mathcal{B} \times_3 Q_i - \mathcal{J}_i||_F^2. \tag{9}$$

According to [7], we perform a model-3 unfolding and singular value decomposition on $\mathcal{S}_i\mathcal{B}$, that is $(\mathcal{S}_i\mathcal{B})^{(3)} = USV^T$. Then let the $\hat{Q}_i = (U(:, 1:d))^T$, where $U(:, 1:d)$ means to select the first d left singular vectors corresponding to the d largest singular values, where the d is the measurement of the intrinsic subspace of the temporal dimension, we empirically set ($d \leq 3$).

4) *Low-rank Approximation* $\mathcal{J}_i$: Dropping the irrelevant variables, we can get following subproblem:

$$\hat{\mathcal{J}}_i = \arg\min_{\mathcal{J}_i} \frac{1}{\lambda_i^2} \parallel \mathcal{S}_i\mathcal{B}\times_3 Q_i - \mathcal{J}_i \parallel_F^2 + ||\mathcal{J}_i||_{tnn}, \tag{10}$$

where λ_i^2 is the noise variance, $\parallel \mathcal{J}_i\parallel_{tnn} = \sum_p \left|\sigma_p(J_i^{(2)})\right|_1$ is the tensor nuclear norm, $J_i^{(2)}$ denotes the model-2 unfolding matrix of the $\mathcal{J}_i$, and $\sigma_p(J_i^{(2)})$ represents the p-th singular value of the $J_i^{(2)}$. This minimization problem is usually solved by the singular value thresholding algorithm [1, 2], we adopt the weighted nuclear norm minimization to boost the performance [3]. Thus, we can get the following formula:

$$\hat{J}_i = U\Psi_W(\mathbf{\Sigma})V^T, \tag{11}$$

where $U\mathbf{\Sigma}V^T$ is the singular value decomposition of the mode-2 unfolding matrix of the $\mathcal{S}_i\mathcal{B}\times_3 Q_i$, and $\Psi_W(\mathbf{\Sigma})_{pp} = \max(\mathbf{\Sigma}_{pp} - W_{pp}, 0)$ is the generalized soft-thresholding operator with the weight vector W. After we obtain the J_i, we conduct the tensor folding to transform it into the 3D tensor $\mathcal{J}_i$.

5) *Clean Video Estimation* $\mathcal{B}$: We fix the other variables and optimize $\mathcal{B}$ with the following subproblem:

$$\min_{\mathcal{B}} \frac{1}{2}||\mathcal{B} + \mathcal{R} - \mathcal{O}\circ\tau||_F^2 + \omega\sum_i \frac{1}{\lambda_i^2}||\mathcal{S}_i\mathcal{B}\times_3 Q_i - \mathcal{J}_i||_F^2 + \gamma||\nabla_t\mathcal{B}||_1, \tag{12}$$

where the first term is the fidelity term, the second term is the low-rank constraint, and the third term represents the local smoothness constraint. The linear operator $\mathcal{S}$ is used to extract the cubic from the video and it can be fast solved on a pixel-by-pixel manner. However, the Eq. (12) is difficult to solve directly. To solve this problem, we introduce the auxiliary variables $\mathcal{Z} = \nabla_t\mathcal{B}$ and $\mathcal{C} = \mathcal{B}$ to spilt the three terms, and the Eq. (12) can be transformed into the following augmented Lagrangian function:

$$\begin{aligned} L_\beta(\mathcal{B},\mathcal{Z},\mathcal{C}) = \frac{1}{2} \parallel \mathcal{B} + \mathcal{R} - \mathcal{O}\circ\tau \parallel_F^2 + \omega\left(\sum_i \frac{1}{\lambda_i^2} \parallel \mathcal{S}_i\mathcal{C}\times_3 Q_i - \mathcal{J}_i \parallel_F^2 + \parallel \mathcal{J}_i\parallel_{tnn}\right) \\ + \gamma \parallel \mathcal{Z}\parallel_1 + \frac{\beta}{2} \parallel \mathcal{Z} - \nabla_t\mathcal{B} - \frac{\Gamma_1}{\beta} \parallel_F^2 + \frac{\beta}{2} \parallel \mathcal{C} - \mathcal{B} - \frac{\Gamma_2}{\beta} \parallel_F^2, \end{aligned} \tag{13}$$

where Γ_1 and Γ_2 are the Lagrangian multipliers, and β is a positive scalar. And the Eq. (13) can be separated into several subproblems:

$$\begin{aligned} \mathcal{B}^{m+1} &= \arg\min_{\mathcal{B}} \frac{1}{2} \parallel \mathcal{B} + \mathcal{R} - \mathcal{O}\circ\tau \parallel_F^2 + \frac{\beta}{2} \parallel \mathcal{Z}^m - \nabla_t\mathcal{B} - \frac{\Gamma_1^m}{\beta} \parallel_F^2 + \frac{\beta}{2} \parallel \mathcal{C}^m - \mathcal{B} - \frac{\Gamma_2^m}{\beta} \parallel_F^2 \\ \mathcal{Z}^{m+1} &= \arg\min_{\mathcal{Z}} \gamma \parallel \mathcal{Z}\parallel_1 + \frac{\beta}{2} \parallel \mathcal{Z} - \nabla_t\mathcal{B}^{m+1} - \frac{\Gamma_1^m}{\beta} \parallel_F^2 \\ \mathcal{C}^{m+1} &= \arg\min_{\mathcal{C}} \omega\sum_i \frac{1}{\lambda_i^2} \parallel \mathcal{S}_i\mathcal{C}\times_3 Q_i - \mathcal{J}_i \parallel_F^2 + \frac{\beta}{2} \parallel \mathcal{C} - \mathcal{B}^{m+1} - \frac{\Gamma_2^m}{\beta} \parallel_F^2 . \end{aligned} \tag{14}$$

We adopt the fast 3-D Fourier transform for fast calculate the clean video $\mathcal{B}$, the following formula is its closed-form solution.

$$\mathcal{B}^{m+1} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathcal{O}\circ\tau - \mathcal{R} + \beta(\nabla_t)^T(\mathcal{Z}^m - \frac{\Gamma_1^m}{\beta})) + \beta(\mathcal{C}^m - \frac{\Gamma_2^m}{\beta}))}{\mathcal{I} + \beta\mathcal{F}^*(\nabla_t)\circ\mathcal{F}(\nabla_t) + \beta\mathcal{I}}\right), \tag{15}$$

where $\mathcal{F}$ is the fast 3-D Fourier transform, $\mathcal{F}^{-1}$ is its inverse transform, and $\mathcal{F}^*$ represents its conjugate transform.

As for the auxiliary variable $\mathcal{Z}$, we can use the soft-thresholding operator to get its closed-form solution:

$$\mathcal{Z}^{m+1} = \Phi_{\frac{\gamma}{\beta}}\left(\nabla_t\mathcal{B}^{m+1} + \frac{\Gamma_1^m}{\beta}\right). \tag{16}$$

For the auxiliary variable $\mathcal{C}$, we have the following formula:

$$\mathcal{C}^{m+1} = \left(\frac{2\omega}{\lambda_i^2}\sum_i (\mathcal{S}_i)^T\mathcal{S}_i + \beta\mathcal{I}\right)^{-1} \times \left(\frac{2\omega}{\lambda_i^2}\sum_i (\mathcal{S}_i)^T\mathcal{J}_i\times_3(Q_i)^T + \beta\mathcal{B}^{m+1} + \Gamma_2^m\right). \tag{17}$$

where $(\mathcal{S}_i)^T\mathcal{J}_i$ is the sum value of all overlapping reconstruction cubics that cover the pixel location, and $(\mathcal{S}_i)^T\mathcal{S}_i$ corresponds to the number of the overlapping cubics. Then the Eq. (17) can be easily solved by the pixel-to-pixel division.

For the Lagrangian multipliers, we update them with the following formulas:

$$\begin{aligned}\Gamma_1^{m+1} &= \Gamma_1^m + \beta\left(\nabla_t \mathcal{B}^{m+1} - \mathcal{Z}^{m+1}\right)\\ \Gamma_2^{m+1} &= \Gamma_2^m + \beta\left(\mathcal{B}^{m+1} - \mathcal{C}^{m+1}\right).\end{aligned} \tag{18}$$

Algorithm 1 The robust low-rank tensor recovery model for acquiring paired GT

Require: The rainy video $\mathcal{O}$

Initialization:;
• Set the regularization parameters μ, ω, λ_i, γ;
• Set the temporal subspace dimension $d \leq 3$;
for l=1:L **do**
  Group similar cubics $\mathcal{S}_i\mathcal{B}^{l+1}$;
  Estimate affine transformation τ^{l+1} via Eq. (5);
  Estimate rain estimation R^{l+1} via Eq. (8);
  Estimate subspace projection ${Q_i}^{l+1}$ via Eq. (9);
  Estimate low-rank approximation ${\mathcal{J}_i}^{l+1}$ via Eq. (11);
  for m=1:M **do**
    Estimate clean video $\mathcal{B}^{m+1}$ via Eq. (15);
    Estimate auxiliary variables $\mathcal{Z}^{m+1}$, $\mathcal{C}^{m+1}$ via Eq. (16) and Eq. (17), respectively;
    Update lagrangian multipliers Γ_1^{m+1} and Γ_2^{m+1} via Eq. (18);
  end for
end for

Ensure: The rain-free video $\mathcal{B}$.

References

[1] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010. 8

[2] Yi Chang, Luxin Yan, and Sheng Zhong. Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising. In *CVPR*, pages 4260–4268, 2017. 6, 8

[3] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *CVPR*, pages 2862–2869, 2014. 8

[4] Huaibo Huang, Mandi Luo, and Ran He. Memory uncertainty learning for real-world single image deraining. *IEEE TPAMI*, 2022. 6

[5] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *NeurIPS*, 24, 2011. 7

[6] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE TPAMI*, 34(11):2233–2246, 2012. 7

[7] Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery. *IEEE TPAMI*, 40(8):1888–1902, 2017. 7

[8] Yuntong Ye, Changfeng Yu, Yi Chang, Lin Zhu, Xi-Le Zhao, Luxin Yan, and Yonghong Tian. Unsupervised deraining: Where contrastive learning meets self-similarity. In *CVPR*, pages 5821–5830, 2022. 3