Self-Learning Video Rain Streak Removal: When Cyclic Consistency Meets Temporal Correspondence

Wenhan Yang[1], Robby T. Tan[2,4], Shiqi Wang[1], Jiaying Liu[3*]
[1] City University of Hong Kong [2] National University of Singapore
[3] Peking University [4] Yale-NUS College

Abstract

In this paper, we address the problem of rain streaks removal in video by developing a self-learned rain streak removal method, which does not require any clean ground-truth images in the training process. The method is inspired by fact that the adjacent frames are highly correlated and can be regarded as different versions of identical scene, and rain streaks are randomly distributed along the temporal dimension. With this in mind, we construct a two-stage Self-Learned Deraining Network (SLDNet) to remove rain streaks based on both temporal correlation and consistency. In the first stage, SLDNet utilizes the temporal correlations and learns to predict the clean version of the current frame based on its adjacent rain video frames. In the second stage, SLDNet enforces the temporal consistency among different frames. It takes both the current rain frame and adjacent rain video frames to recover the structural details. The first stage is responsible for reconstructing main structures, and the second stage is responsible for extracting structural details. We build our network architecture with two sub-tasks, i.e. motion estimation and rain region detection, and optimize them jointly. Our extensive experiments demonstrate the effectiveness of our method, offering better results both quantitatively and qualitatively.

1. Introduction

Rain is a common bad weather condition that introduces a series of visibility degradation in captured videos and images. The presence of rain not only leads to poor visual quality but also impairs existing computer vision systems that assume clean video frames as input. Rain streaks are the most common type of rain degradation. They can partially occlude a background scene, change image appearance, make the scene blurred, *etc*. Besides rain streaks, rain also generates a veiling effect (visually similar to fog) and raindrops that are attached to a lens or windscreen. In our paper, we focus on rain streaks removal. Our method learns from rain videos themselves, without requiring any ground-truth clean videos in the training process.

(a) Rain Frame (b) FastDeRain [22] (c) MS-CSC [28] (d) Ours

Figure 1. Visual results of different deraining methods on a real rain video frame with large motions. Compared to MS-CSC [28] and FastDeRain [22], our self-learned method does not have artifacts and is more effective in removing rain streaks. Note that, SpacCNN is a fully-supervised method, while ours is self-learned, which does not require any clean ground-truth videos in the training phase.

A few existing methods [24, 20, 38, 34] focus on separating rain-free background images (clean images) and rain streaks based on spatial redundancies and detail texture appearances. Several mathematical models extract discriminative features to set part these two layers, *e.g.* the frequency domain representation [24], sparse representation [34], Gaussian mixture models [29] and deep networks [47, 14].

Beyond only exploiting spatial redundancies, video-based methods [1, 2, 3, 9, 12, 15, 17, 18, 53] further utilize temporal correlations and contexts to address the problem. The earliest methods [17, 15, 18] leverage the physical and photometric properties, *i.e.* the directional and chromatic

*Corresponding author. Email: liujiaying@pku.edu.cn. This work is partially supported by National Natural Science Foundation of China under contract No.61772043, in part by Beijing Natural Science Foundation under contract No.L182002, in part by the National Key R&D Program of China under Grand No.2018AAA0102700, and in part by the Hong Kong ITF UICP under Grant 9440203. Robby T. Tan's research in this work is supported by MOE2019-T2-1-130.

properties of rain streaks. Later approaches [9, 6, 26, 23] further make use of the temporal dynamics of videos, i.e. the continuity of background layers and the randomness of rain locations along the temporal dimension, to remove rain streaks from rain videos. The efficiency of deep learning also leads to the emergence of deep-learning based video rain streak removal [28, 32, 30, 45, 7]. Convolutional neural network (CNN) as well as other advanced deep models are developed to better separate rain streaks and background scene, *e.g.* recurrent neural network [32, 30], convolutional sparse coding [28]. Many effective priors and features are designed, *e.g.* explicit temporal correlations [7], scale varieties of rain streaks [28] and motion contexts [32, 30].

With the power of learning from data, the CNN-based methods outperform previous traditional methods. However, most of these learning methods require ground-truth images that are free from rain streaks. For this, they employ synthetic rain images, since to have pairs of real rain images and ground-truth rain-free images is intractable to obtain. Hence, the accuracy of these methods depends on the quality of the synthesized rain.

In this paper, we aim to develop a deep-learning based method that does not require any clean video ground-truths in the training process. Thus, unlike most of the learning methods, we do not use any rain synthetic training data. By making full use of temporal correlation and consistency, a two-stage **S**elf-**L**earned **D**eraining **Net**work (**SLDNet**) is designed to learn how to remove rain streaks solely from the input videos and some priors. The first stage of SLDNet predicts the clean current frame based on its adjacent rain video frames via video frame interpolation, without any information of the current rain frame. As the rain streaks distribute randomly along the temporal dimension, the result of this stage is almost rain-free. Yet, when large motions are present, some details are blurred due to the intrinsic difficulty in modeling large/fast motions. To avoid these artifacts, in the second stage of SLDNet, we include the information of the current rain frame, bringing in some texture details while filtering out the rain streaks. Our temporal consistency constraint forces the generated result (with the added texture details) to be close to other adjacent aligned rain video frames. Along the way, some motion prior and rain-related prior are injected into our method.

In summary, our contributions are as follows.

- We propose a self-learned video rain streak removal method that can learn solely from input videos. To our knowledge, this is the first attempt in video-based rain streak removal literature. Integrated with both temporal correlation and consistency, the proposed deep network, first, infers the main structures of a clean frame, and then recovers the details.

- Besides the temporal correlation and consistency constraint, we further inject priors of rain videos, *i.e.* background motion and rain location information, to benefit rain streaks removal without requiring any paired rain-free ground-truths in the network training. These constraints/priors can possibly open up further the exploration of self-learned video rain streak removal.

- We propose a framework that jointly optimizes the background motion and rain localization while removing rain streaks. Extensive experiments demonstrate the effectiveness of our joint optimization and thus the effectiveness of our whole method.

2. Related Work

As the rain causes poor visibility, occludes the background scene and blurs the background, rain removal methods are proposed to restore the clean image from a rain one. One branch is the single-image rain streak removal, which aims to infer the clean image solely based on a single rain image. Many models are developed to capture the intrinsic differences between the rain signal and normal textures based on the spatial redundancy, *e.g.* generalized low rank model [9], sparse coding [24], discriminative sparse coding [34], nonlocal mean filter [25], Gaussian mixture model [29], transformed low rank model [4], rain direction prior [51]. In 2017, single-image deraining steps into the era of deep-learning and many deep-learning based methods emerge, including deep detail network [14, 13], joint rain detection and removal [47, 48], density-aware multi-stream densely connected CNN [51], perceptual generative adversarial network [41]. Later works focus on developing advanced deep networks [27, 42, 35, 49] or utilizing more effective priors [19, 5, 54, 50, 52, 46].

Compared with single-image rain removal, video rain streak removal is capable of utilizing temporal correlation and dynamics to detect and remove rains. Garg and Nayar propose the seminal work of video rain modeling [17] and rain streak removal methods [15, 18, 16]. Later approaches dig deep to see the intrinsic priors rain streak and normal background signals, *i.e.* temporal and chromatic properties of rain [53, 33], the size, shape and orientation of rain streaks [3, 2], phase congruency features [37], Fourier domain feature [1], spatio-temporal correlation of patch groups [9], rain directional prior of rain streaks [23], Gaussian mixture model [6], Bayes rain detector [39, 40], two-stage detection and refinement based on SVM [26], patch-based mixtures of Gaussian[44], matrix decomposition [36]. Recently, deep-learning based methods bring significant changes to video deraining with augmented capacities and flexibilities. In [28], Li *et al.* apply a multiscale convolutional sparse coding to remove the rain streaks with different scales. Chen *et al.* [8] propose to firstly segment superpixels from a rain frame and then to estimate rain-free superpixels with the consistency constraint among the

aligned super-pixels. After that, compensate lost details, a CNN is further used to add normal textures to the final results. In [31], Liu *et al.* build a recurrent neural network that seamlessly integrates rain degradation classification, rain removal and background details reconstruction. In [32], a hybrid rain model is proposed to model both rain streaks and occlusions, and is then injected into a dynamic routing residue recurrent network with the motion segmentation context information. In [45], a two-stage recurrent network is built with dual-level flow regularizations to perform the inverse recovery process of the rain synthesis model for video deraining.

Previous works are either model-based, designed with hand-crafted features, or data-driven ones, relying on synthetic paired data. In our work, we explore the possible architectures and priors for self-learning and construct a learnable video deraining network which does not rely on synthesized paired data.

3. Rain Modeling and Self-Learning Constraint

3.1. Rain Video Modeling

We formulate a rain model as:

$$I = B + R, \tag{1}$$

where B is the layer without rain streaks, and R is the rain streak layer. I is the captured image with rain streaks. A video rain synthesis model is obtained with a temporal indicator t added:

$$I_t = B_t + R_t,\ t = 1, 2, ..., N, \tag{2}$$

where t and N denote the current time-step and the total number of video frames, respectively. The rain streak R_t is assumed to be independent and identically distributed random samples. There are also more complicated rain synthesis models, *e.g.* [45] that take into account the rain accumulation, flow, *etc*. In this paper, we only consider the problem of rain streak removal by exploring the information from rain videos.

3.2. Temporal Cyclic Consistency for Self-Learned Rain Removal

We explore intrinsic constraints and priors that facilitate video rain streak removal even without paired training data, specifically our constraints/priors are consisting of three aspects: temporal correlation, temporal consistency, and rain-related priors.

Temporal Correlation. Adjacent clean video frames are highly correlated. Meaning, the background signal of a rain frame can be predicted by its adjacent rain video frames, since rain streaks are likely randomly distributed. Therefore, *if we try to predict a current rain frame based on adjacent rain video frames (without the current one), the rain signal will not be predicted and the result will tend to be rain-free.* However, when the frames include large motions, it is also challenging to interpolate a frame based on its adjacent frames, which can lead to blurred details and artifacts.

Temporal Consistency. Because non-rain background layers are continuous along the temporal dimension, the video frames after motion compensation should be well aligned and lead to small differences. Comparatively, even good motion estimation and compensation are achieved, the well aligned rain layers are also very different, due to the existence of rain streaks. Hence, *it is beneficial to remove rain streaks if we enforce the model to generate the consistent results after motion compensation*. However, motion might not be well estimated if large motions are present, and there may be content changes among different frames. In this case, the temporal consistency regularization might also fail. Therefore, in our work, we also include the motion estimation as part of our optimization target.

Rain-Related Side Information. Besides the above two constraints to connect rain video frame and their corresponding rain-free versions, we also intend to embed useful side information to guide the deraining process. The rain-dependent features, *i.e.* rain mask, can be injected as a part of the loss functions, which control the model to process rain layers adaptively, namely only applying rain removal in the rain regions. Another kind of features, *i.e.* optical flow, whose estimation is usually extracted from clean frames, are easy to be contaminated by the appearance of rain. Optical flow estimation has a complicated and intertwined effect on the rain streak removal. However, optical flow and rain removal can benefit each other if one of their performance is improved. Hence, optical flow estimation is regarded as one part of our whole optimization function.

4. Self-Learned Deraining Network

4.1. Network Architecture

Based on the discussion in the last section, we build a **S**elf-**L**earned **D**eraining **Net**work (**SLDNet**) as shown in Fig. 2, which consists of three parts:

- Warping operation (Fig. 2 (a)). This part extracts the optical flow [11] as the motion information and apply alignment among frames. This module (particularly optical flow) is jointly optimized with the whole deraining task.
- Prediction Network (PredNet) (Fig. 2 (b)). In the training phase, the network aims to predict the rain-free background layer of the current frame, based on its adjacent rain video frames.
- EnHancement Network (EHNet) (Fig. 2 (c)). Guided

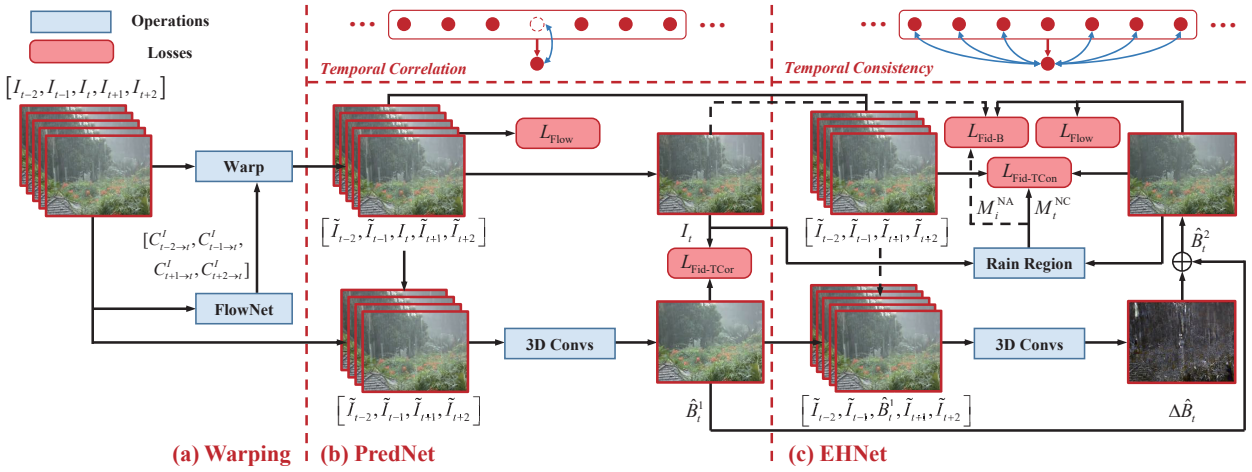


Figure 2. The framework of our proposed Self-Learning Deraining Network (SLDNet). 1) *Warping* module aligns the the neighboring frames to the central one. Successive modules make full use of *temporal correlations* and *consistency* to create the mapping from rain video frames to the clean ones. 2) *Prediction Network* (*PredNet*) predicts the clean version of the current frame with the neighboring rain video frames, taking the rain version of the current one as the ground truth. 3) *EnHancement Network* (*EHNet*) compensates for the detail of the predicted clean layers with both the neighboring and current rain video frames, taking the aligned version of the neighboring rain video frames as the ground truth. The red arrows denotes the direction of the information flow and the blue arrows denote main related constraints and losses.

by the rain-free estimation produced by PredNet, we then improve the details via an enhancement network. The network takes the current rain frame and the adjacent rain video frames as input, and generates the residual details under the inter-frame consistency constraint. A rain mask is incorporated into the loss function to reduce the impact of rain streaks in the ground truth on the deraining process, and make the network only focus on the useful information in non-rain regions.

In the subsequent sections, we discuss each part of the network in details.

4.2. Proposed Networks

1) Optical Flow Estimation and Warping (Fig. 2 (a)). We first estimate the optical flow and warp the input rain video frames. $G(\cdot)$ is introduced to denote the processes to extract optical flow from the given image pair as follows,

$$C^I_{i\to j} = G(I_i, I_j), \tag{3}$$

$$C^B_{i\to j} = G(\hat{B}_i, \hat{B}_j), \tag{4}$$

where the subscript $i \to j$ denotes the flow from the i-th frame to the j-th one, and superscript I and B denote the flow is estimated from the rain image or the estimated background image. Then, we can warp the image to the j-th time-step based on the estimated flow:

$$\widetilde{I}^I_{i\to j} = W(I_i, C^I_{i\to j}), \tag{5}$$

$$\widetilde{B}^B_{i\to j} = W(\hat{B}_i, C^B_{i\to j}). \tag{6}$$

For simplicity, in Fig. 2, we use $\widetilde{I}^I_i$ to denote $\widetilde{I}^I_{i\to j}$ as j is set to t for the whole process. To improve flow estimation accuracy, in the training phase, we finetune the pretrained optical flow network with the rain video frames and the estimated background layers. After the warping, these frames should be well aligned to the current ones, expressed as:

$$\begin{aligned}\mathcal{L}_{\text{Flow}} = &\sum_{i=t-s}^{t-1} \left\| \widetilde{I}^I_{i\to t} - I_t \right\|_2^2 + \sum_{i=t+1}^{t+s} \left\| \widetilde{I}^I_{i\to t} - I_t \right\|_2^2 \\ &+ \sum_{i=t-s}^{t-1} \left\| \widetilde{B}^I_{i\to t} - \hat{B}^R_t \right\|_2^2 + \sum_{i=t+1}^{t+s} \left\| \widetilde{B}^I_{i\to t} - \hat{B}^R_t \right\|_2^2. \end{aligned} \tag{7}$$

When the background layers are recovered, they provide more accurate information to learn better estimation of optical flow.

2) PredNet (Fig. 2 (b)). We use $F_{\text{Interp}}(\cdot)$ to denote the rain frame interpolation process, where the current rain frame I_t is not involved in the input. The model is trained based on *temporal correlation*, where the information is flowed from adjacent rain video frames to the current rain one. Then, the initial rain-free result $\hat{B}^1_t$ is predicted as follows,

$$\hat{B}^1_t = F_{\text{Interp}}\left(\widetilde{\Phi}^I_{t,s} / \{I_t\}\right), \tag{8}$$

$$\widetilde{\Phi}^I_{t,s} = \left\{ \widetilde{I}^I_{(t-s)\to t}, ..., \widetilde{I}^I_{(t-1)\to t}, I^I_t, \widetilde{I}^I_{(t+1)\to t}, ..., \widetilde{I}^I_{(t+s)\to t} \right\}.$$

The loss function is defined as follows,

$$\mathcal{L}_{\text{Fid-TCor}} = \left\| \hat{B}^1_t - I_t \right\|_2^2. \tag{9}$$

3) EHNet (Fig. 2 (c)). We use $F_{\text{Enhance}}(\cdot)$ to denote the detail enhancement process, where the current rain frame and adjacent frames $\widetilde{\Phi}^I_{t,s}$ are all taken as the input. The model is trained based on *temporal consistency*, requiring the estimated frame to be consistent to all of the aligned adjacent rain rain video frames.

$$\hat{B}^2_t = F_{\text{Enhance}}\left(\widetilde{\Phi}^I_{t,s}\right) + \hat{B}^1_t. \tag{10}$$

The loss function is defined as follows,

$$\mathcal{L}_{\text{Fid-TCon}} = \sum_{i=\{t-s,\dots,t+s\}/t} \frac{1}{2s} \left\| M^{\text{NA}}_i \left(\widetilde{B}^2_i - I_i\right)\right\|^2_2, \tag{11}$$

$$\widetilde{B}^2_i = W\left(\hat{B}^2_t, C^I_{t\to i}\right), \tag{12}$$

where M^{NA}_i is the estimated mask of the non-rain region of the adjacent rain frame I_i, which will be further discussed in the following paragraphs.

4) Rain Region Estimation. Having gone through enough training time, $\hat{B}^2_t$ will be accurate enough. Then, we can infer the non-rain region M^{NC}_t of the current frame, a soft mask denoting whether pixels are not covered by rain streaks, and the non-rain regions M^{NA}_t of the adjacent rain video frames, also soft masks denoting whether a pixel is free of rain streaks. M^{NC}_t is calculated as follows,

$$M^{\text{NC}}_t = \exp\left\{-\frac{\left(g_{\text{ReLU}}\left(I_t - \hat{B}^2_t\right)\right)^2}{\omega}\right\}, \tag{13}$$

where ω controls the shape of the exponential function in Eq. (13). $g_{\text{ReLU}}(\cdot)$ is the rectified linear unit function that gets only the positive values passed, which is decided by the common observation of positive rain streaks.

As for Eq. (11), we find that, if there is no mask involved, the effectiveness of the loss largely depends on accuracy of the motion estimation, and the degree of the true intrinsic inter-frame correspondence. When the reliable rain masks can be acquired, the guidance of the loss can be augmented in two ways. First, as denoted in Eq. (11), it makes the model learn useful information only from non-rain regions of adjacent frames. Second, we can also guide the model to learn more from the non-rain regions of the current rain frame. With this in mind, we augment the whole fidelity loss in the enhancement stage as follows,

$$\begin{aligned}\mathcal{L}_{\text{Fid-T}} &= \mathcal{L}_{\text{Fid-TCon}} + \mathcal{L}_{\text{Fid-TCor}},\\ \mathcal{L}_{\text{Fid-B}} &= M^{\text{NC}}_t \left\|\hat{B}^2_t - I_t\right\|^2_2.\end{aligned} \tag{14}$$

The injection of the rain masks regularizes the model to process different regions adaptively. For rain regions, the output tends to be consistent to the corresponding regions of the aligned rain video frames. For non-rain regions, the output preserves the information of the input rain frame.

5) Overall Loss Function. The whole loss function is the summation of the above mentioned losses:

$$\mathcal{L}_{\text{All}} = \mathcal{L}_{\text{Flow}} + \lambda_{\text{T}}\mathcal{L}_{\text{Fid-T}} + \lambda_{\text{B}}\mathcal{L}_{\text{Fid-B}}, \tag{15}$$

where λ_{T} and λ_{B} are two weighting parameters that balance the importance of each term. This loss will guide the network to learn to remove rain streaks from input rain videos.

6) Discussion on Joint Optimization of Rain-Related Priors. Our framework considers to inject rain-related priors into our framework from the following aspects:

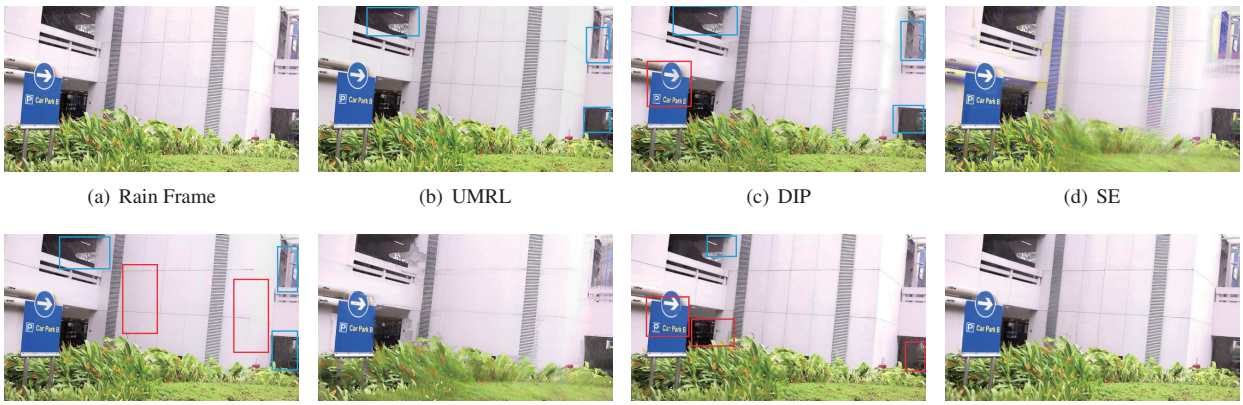
- Optical flow finetuning. Optical flow estimates the pixel-level motion for frame alignment and warping. When it is estimated more accurately, the performance of rain removal and rain region estimation will significantly improve. Inversely, when more accurate deraining results are achieved, the optical flow estimation can also improve through Eq. (7).

- Rain region estimation. More accurate deraining results and optical flow estimation lead to better rain region estimation via Eq. (13). Also, better estimated rain regions bring in more accurate deraining results through Eq. (14).

5. Experimental Results

Datasets. We compare our model with state-of-the-arts on *NTURain* [8], which has two sub-groups: one taken from a panning and unstable camera with slow movements, and the other from a fast moving car-mount camera. There are also other video rain datasets, *e.g.* *RainSynLight25* and *RainSynComplex25* [31]. However, the frame length of videos in these datasets is too short (only 7-15 frames) to perform self-training. Thus, we do not compare different methods on these datasets. We also compare several real rain videos generally used in the previous methods and those from Youtube as well as our own rain data. More visual results, *e.g.* optical flow, rain mask estimation, and video results are provided in the supplementary material.

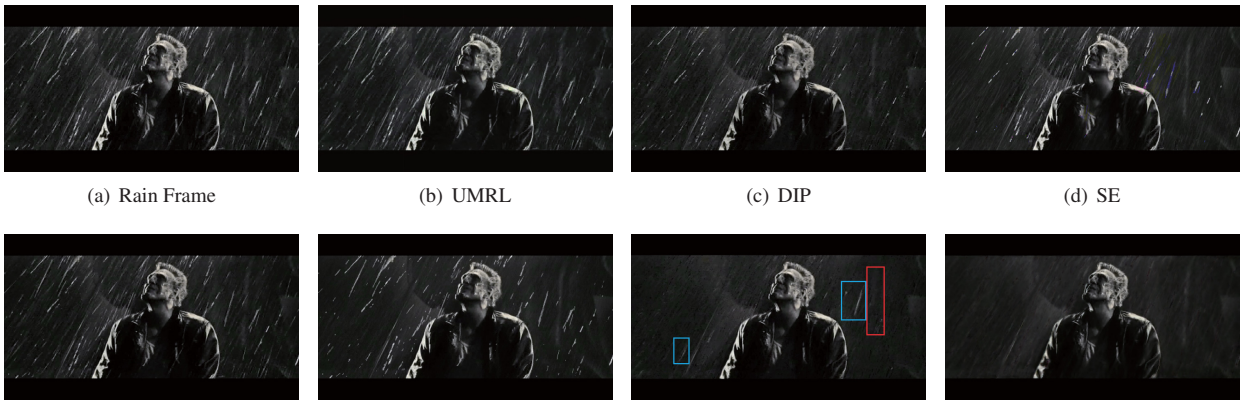
Implementation Details. We use *NTURain* and our collected real rain videos for evaluation. *NTURain* includes 25 paired videos for training and 8 for testing. However, for the quantitative evaluation, we do not use *NTURain*'s training set at all, and use only its testing set (as our method can self-learn). For our qualitative evaluation, we use the collected real rain videos that do not have the paired clean version. Our deraining networks (PredNet and EHNet) are trained using Adam optimizer with the learning rate $1e^{-4}$.



(a) Rain Frame (b) UMRL (c) DIP (d) SE

(e) FastDeRain (f) MS-CSC (g) SpacCNN (h) Ours

Figure 3. Visual comparison of different deraining methods on *rb3* in *NTURain*. The remaining rain streaks and artifacts are denoted with blue and red boxes, respectively.



(a) Rain Frame (b) UMRL (c) DIP (d) SE

(e) FastDeRain (f) MS-CSC (g) SpacCNN (h) Ours

Figure 4. Visual comparison of different deraining methods on a real rain video sequence. The remaining rain streaks and artifacts are denoted with blue and red boxes, respectively.

The optical flow module comes with its existing pretrained model and is finetuned to our input with the learning rate $1e^{-7}$. All training videos are sampled and cropped into $64\times64\times5$ cubics with a batch size of 8. The PSNR and SSIM results are the average results of all frames of the input sequence. As for the comparison settings, the results of all compared methods are generated using the authors' provided codes and settings: MS-CSC and J4RNet are trained with their own datasets, URML and PReNet are trained with *Rain100H*, and other methods are traditional methods, which do not rely on the training data. We consider not training these networks using *NTURain*'s training set is fair, since: (1) practically, in the deployment stage, most of the time we do not know the domain of the videos, (2) our network is also not trained using *NTURain*'s training set.

Baselines. We compare the proposed network with state-of-the-art methods: Uncertainty guided Multi-scale Residual Learning (UMRL) [50], Directional Global Sparse Model (UGSM) [10], Progressive Recurrent Network (PReNet) [35], Discriminatively Intrinsic Priors (DIP) [23], FastDeRain [22], Stochastic Encoding (SE) [44], Multi-Scale Convolutional Sparse Coding (MS-CSC) [28], Joint Recurrent Rain Removal and Reconstruction Network (J4RNet) [31], SuperPixel Alignment and Compensation CNN (SpacCNN) [8]. UMRL, UGSM, and PReNet are single frame deraining methods offering state-of-the-art performance in single-image rain removal. SE, DIP, FastDerain, J4RNet, MS-CSC, and SpacCNN are multi-frame derainig methods. UMRL, PReNet, J4RNet and SpacCNN are deep-learning based methods. All methods are tested with the codes kindly released by the authors. For the experiments on synthesized data, Peak Signal-to-Noise Ratio (PSNR) [21] and Structure Similarity Index (SSIM) [43] are used as comparison criteria. Following previous works, we evaluate the results only in the luminance channel, since human visual system is more sensitive to lu-

(a) Rain Frame (b) UMRL (c) DIP (d) SE

(e) FastDeRain (f) J4R (g) SpacCNN (h) Ours

Figure 5. Visual comparison of different deraining methods on a real rain video sequence. The remaining rain streaks and artifacts are denoted with blue and red boxes, respectively.

(a) Rain Frame (b) UMRL (c) PReNet (d) SE

(e) MS-CSC (f) FastDerain (g) SpacCNN (h) Ours

Figure 6. Visual comparison of different deraining methods on a real rain video sequence. The remaining rain streaks and artifacts are denoted with blue and red boxes, respectively.

minance than chrominance information.

Quantitative Evaluation. We first compare the performance of different methods quantitatively in Table 1. Comparing different methods, including both single-image deraining methods and multi-frame rain removal methods, several observations are obtained. First, our results are consistently better than previous methods, either data-driven approaches or low-rank based methods, which further shows the subtle of our model design. Second, compared with the state-of-the-art single-image deraining method, URML, PReNet and UGSM, our method achieves at least 3dB and 0.01 gain in PSNR and SSIM, respectively, which show the importance of temporal modeling beyond the big data knowledge. Third, our method significantly outperforms SpacCNN, the state-of-the-art deep-learning video deraining method, with a gain of 1.6 dB 0.0065 in PSNR and SSIM, respectively.

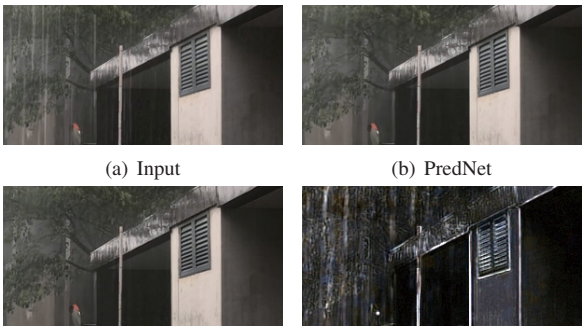
Qualitative Evaluation. We also compare results of differ-

Table 1. PSNR and SSIM results among different rain streak removal methods on *NTURain*. Best results are denoted in red and the second best results are denoted in blue.

Metric	Rain	URML	PReNet	UGSM	MS-CSC	DIP	SE	FastDeRain	J4RNet	SpacCNN	Ours
PSNR	30.41	31.33	30.36	31.29	26.64	30.79	26.04	30.54	30.73	33.11	34.89
SSIM	0.9108	0.9477	0.9437	0.9253	0.7661	0.9370	0.7571	0.9255	0.9407	0.9475	0.9540

Table 2. Ablation study for the two-stage network architecture on *NTURain*. Best results are denoted in bold.

Metric	PredNet	EHNet	PredNet+EHNet
PSNR	33.61	33.62	**34.89**
SSIM	0.9436	0.9465	**0.9537**



(a) Input (b) PredNet

(c) PredNet + EHNet (d) Difference Map

Figure 7. Visualization of stage-wise results of our method. The difference map shows how EHNet compensates for the details based on the result of PredNet.

(a) Rain Frame (b) DIP (c) SpacCNN (d) Ours

Figure 8. Limitations of existing methods.

Table 3. Ablation study for rain-related priors used in our work on a sub-set of *NTURain* (*b1-b4*). SLDNet-v1: without optical flow finetuning (OFF), rain region guidance (RRG). SLDNet-v2: with OFF but without RRG. SLDNet-full: with OFF, RRG. Best results are denoted in bold.

Metric	SLDNet-v1	SLDNet-v2	SLDNet-full
PSNR	34.35	35.48	35.78
SSIM	0.9546	0.9577	0.9588

ent methods qualitatively. Three groups of results on real videos and one group of results on *NTURain* are provided in Figs. 3-6. The testing videos include diversified kinds of rain streaks in scale, density and intensity. Our results provide more effective results, with less remaining rain streaks, abundant details, and less blurring and artifacts. More visual results will be provided in the supplements.

Ablation Study of Two-Stage Network Architecture. To demonstrate the effectiveness of our two-stage network design, we perform an ablation study on the network architecture in Table 2. It is observed that, the combination of PredNet and EHNet achieves more effective performance to any single one, which confirms necessity of utilizing temporal correlation and consistency jointly. We also provide visual results in Fig. 7. It is clearly shown that, EHNet effectively further suppresses residual rain streaks and enhances the details.

Ablation Study of Rain-Related Priors. We also compare different versions of our deraining network with and without rain priors. The involved priors include optical flow and rain region mask. The results are shown in Table. 3. Comparing SLDNet-v1, and SLDNet-v2, it is observed that, optical flow finetuning contributes to a large performance gain. Furthermore, comparing SLDNet-v2 and SLDNet-full, it is showed that, it is also beneficial to inject the rain region mask, which tells the network which ground truths are more reliable.

Limitation and Future Direction. Although achieving impressive results in many cases, when the video includes large and irregular motions, our method might generate blurred results. One example is shown in Fig. 8. Our results might be more effective than the results of DIP and SpacCNN. However, our result is still not promising. In the future, it might be expected to adopt adversarial learning to model natural background layers to better preserve normal textures of the results.

6. Conclusion

In this paper, we make the first attempt to address the problem of the video rain streak removal with only the information of rain video frames. A Self-Learned Deraining Network (SLDNet) is built to make full use of both temporal correlation and consistency to obtain the mapping between the rain video frames and clean ones. The network is a two-stage architecture. In the first stage, the model learns to predict the clean current frame based on its adjacent rain video frames (without the current rain frame), taking the current rain frame as the ground truth. The second stage takes both the current rain video frame and adjacent rain video frames for detail compensation, where the result is constrained to be close to aligned adjacent rain video frames. The first stage reconstructs main structures and the second stage compensates for structural details. The effective rain-related priors are injected to the model. The extensive experiments demonstrate the effectiveness of our method, its each component, and the related rain priors.

References

[1] Peter C Barnum, Srinivasa Narasimhan, and Takeo Kanade. Analysis of rain and snow in frequency space. *Int'l Journal of Computer Vision*, 86(2-3):256–274, 2010. 1, 2

[2] Jérémie Bossu, Nicolas Hautière, and Jean-Philippe Tarel. Rain or snow detection in image sequences through use of a histogram of orientation of streaks. *International journal of computer vision*, 93(3):348–367, 2011. 1, 2

[3] Nathan Brewer and Nianjun Liu. Using the shape characteristics of rain to identify and remove rain from video. In *Joint IAPR International Workshops on SPR and SSPR*, pages 451–458, 2008. 1, 2

[4] Yi Chang, Luxin Yan, and Sheng Zhong. Transformed low-rank model for line pattern noise removal. In *Proc. IEEE Int'l Conf. Computer Vision*, Oct 2017. 2

[5] Y. Chang, L. Yan, and S. Zhong. Transformed low-rank model for line pattern noise removal. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1735–1743, Oct 2017. 2

[6] J. Chen and L. P. Chau. A rain pixel recovery algorithm for videos with highly dynamic scenes. *IEEE Trans. on Image Processing*, 23(3):1097–1104, March 2014. 2

[7] J. Chen, C. Tan, J. Hou, L. Chau, and H. Li. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 6286–6295, June 2018. 2

[8] Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2018. 2, 5, 6

[9] Yi-Lei Chen and Chiou-Ting Hsu. A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1968–1975, 2013. 1, 2

[10] Liang-Jian Deng, Ting-Zhu Huang, Xi-Le Zhao, and Tai-Xiang Jiang. A directional global sparse model for single image rain removal. *Applied Mathematical Modelling*, 59:662 – 679, 2018. 6

[11] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE Int'l Conf. Computer Vision*, 2015. 3

[12] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proc. IEEE Int'l Conf. Computer Vision*, December 2013. 1

[13] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Trans. on Image Processing*, 26(6):2944–2956, June 2017. 2

[14] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, July 2017. 1, 2

[15] Kshitiz Garg and Shree K Nayar. Detection and removal of rain from videos. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, volume 1, pages I–528, 2004. 1, 2

[16] Kshitiz Garg and Shree K Nayar. When does a camera see rain? In *Proc. IEEE Int'l Conf. Computer Vision*, volume 2, pages 1067–1074, 2005. 2

[17] Kshitiz Garg and Shree K Nayar. Photorealistic rendering of rain streaks. In *ACM Trans. Graphics*, volume 25, pages 996–1002, 2006. 1, 2

[18] Kshitiz Garg and Shree K Nayar. Vision and rain. *Int'l Journal of Computer Vision*, 75(1):3–27, 2007. 1, 2

[19] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2019. 2

[20] De-An Huang, Li-Wei Kang, Yu-Chiang Frank Wang, and Chia-Wen Lin. Self-learning based image decomposition with applications to single image denoising. *IEEE Transactions on multimedia*, 16(1):83–93, 2014. 1

[21] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6

[22] T. Jiang, T. Huang, X. Zhao, L. Deng, and Y. Wang. Fastderain: A novel video rain streak removal method using directional gradient priors. *IEEE Trans. on Image Processing*, 28(4):2089–2102, April 2019. 1, 6

[23] Tai-Xiang Jiang, Ting-Zhu Huang, Xi-Le Zhao, Liang-Jian Deng, and Yao Wang. A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, July 2017. 2, 6

[24] L. W. Kang, C. W. Lin, and Y. H. Fu. Automatic single-image-based rain streaks removal via image decomposition. *IEEE Trans. on Image Processing*, 21(4):1742–1755, April 2012. 1, 2

[25] J. H. Kim, C. Lee, J. Y. Sim, and C. S. Kim. Single-image deraining using an adaptive nonlocal means filter. In *Proc. IEEE Int'l Conf. Image Processing*, pages 914–917, Sept 2013. 2

[26] J. H. Kim, J. Y. Sim, and C. S. Kim. Video deraining and desnowing using temporal correlation and low-rank matrix completion. *IEEE Trans. on Image Processing*, 24(9):2658–2670, Sept 2015. 2

[27] Guanbin Li, Xiang He, Wei Zhang, Huiyou Chang, Le Dong, and Liang Lin. Non-locally enhanced encoder-decoder network for single image de-raining. In *ACM Trans. Multimedia*, pages 1056–1064. ACM, 2018. 2

[28] Minghan Li, Qi Xie, Qian Zhao, Wei Wei, Shuhang Gu, Jing Tao, and Deyu Meng. Video rain streak removal by multiscale convolutional sparse coding. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2018. 1, 2, 6

[29] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2736–2744, 2016. 1, 2

[30] J. Liu, W. Yang, S. Yang, and Z. Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 3233–3242, June 2018. 2

[31] Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2018. 3, 5, 6

[32] J. Liu, W. Yang, S. Yang, and Z. Guo. D3r-net: Dynamic routing residue recurrent network for video rain removal. *IEEE Trans. on Image Processing*, 28(2):699–712, Feb 2019. 2, 3

[33] Peng Liu, Jing Xu, Jiafeng Liu, and Xianglong Tang. Pixel based temporal analysis using chromatic property for removing rain from videos. In *Computer and Information Science*, 2009. 2

[34] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 3397–3405, 2015. 1, 2

[35] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2019. 2, 6

[36] Weihong Ren, Jiandong Tian, Zhi Han, Antoni Chan, and Yandong Tang. Video desnowing and deraining based on matrix decomposition. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, July 2017. 2

[37] Varun Santhaseelan and Vijayan K. Asari. Utilizing local phase information to remove rain from video. *Int'l Journal of Computer Vision*, 112(1):71–89, March 2015. 2

[38] Shao-Hua Sun, Shang-Pu Fan, and Yu-Chiang Frank Wang. Exploiting image structural similarity for single image rain removal. In *Proc. IEEE Int'l Conf. Image Processing*, pages 4482–4486, 2014. 1

[39] Abhishek Kumar Tripathi and Sudipta Mukhopadhyay. A probabilistic approach for detection and removal of rain from videos. *IETE Journal of Research*, 57(1):82–91, 2011. 2

[40] A. K. Tripathi and S. Mukhopadhyay. Video post processing: low-latency spatiotemporal approach for detection and removal of rain. *IET Image Processing*, 6(2):181–196, March 2012. 2

[41] C. Wang, C. Xu, C. Wang, and D. Tao. Perceptual adversarial networks for image-to-image transformation. *IEEE Trans. on Image Processing*, 27(8):4066–4079, Aug 2018. 2

[42] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson W.H. Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2019. 2

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004. 6

[44] Wei Wei, Lixuan Yi, Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Should we encode rain streaks in video as deterministic or stochastic? In *Proc. IEEE Int'l Conf. Computer Vision*, Oct 2017. 2, 6

[45] Wenhan Yang, Jiaying Liu, and Jiashi Feng. Frame-consistent recurrent video deraining with dual-level flow. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2019. 2, 3

[46] W. Yang, J. Liu, S. Yang, and Z. Guo. Scale-free single image deraining via visibility-enhanced recurrent wavelet learning. *IEEE Trans. on Image Processing*, 28(6):2948–2961, June 2019. 2

[47] 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 *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, July 2017. 1, 2

[48] W. Yang, R. T. Tan, J. Feng, J. Liu, S. Yan, and Z. Guo. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 2

[49] Wenhan Yang, Shiqi Wang, Dejia Xu, Xiaodong Wang, and Jiaying Liu. Towards scale-free rain streak removal via self-supervised fractal band learning. In *Proc. AAAI Conf. on Artificial Intelligence*, Feb. 2020. 2

[50] Rajeev Yasarla and Vishal M. Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2019. 2, 6

[51] He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2018. 2

[52] He Zhang, Vishwanath Sindagi, and Vishal M. Patel. Image De-raining Using a Conditional Generative Adversarial Network. *arXiv e-prints*, page arXiv:1701.05957, Jan 2017. 2

[53] Xiaopeng Zhang, Hao Li, Yingyi Qi, Wee Kheng Leow, and Teck Khim Ng. Rain removal in video by combining temporal and chromatic properties. In *Proc. IEEE Int'l Conf. Multimedia and Expo*, pages 461–464, 2006. 1, 2

[54] L. Zhu, C. Fu, D. Lischinski, and P. Heng. Joint bi-layer optimization for single-image rain streak removal. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 2545–2553, Oct 2017. 2