Self-Supervised Architecture for Online Deraining

Aditya Mohan

Abstract

Rain significantly degrades image quality, posing challenges for vision-based tasks. Traditional deraining methods rely on supervised learning with paired datasets, which are difficult to compile due to the dynamic nature of rain. Synthetic datasets, while a common alternative, often lack realism and fail to capture the diversity of real-world scenarios. We propose a self-supervised method for image deraining that leverages the spatial movement of rain across frames relative to static scene elements. By using depth and camera pose information, we align scenes and apply a view-synthesis constraint to generate pseudo-ground truth images, isolating clean pixels from warped frames. These pseudo-clean images enable effective rain removal without requiring paired datasets. Evaluations on real-world rainy datasets show that our method outperforms state-of-the-art unsupervised approaches, achieving superior deraining performance.

1 Introduction

Rain adversely impacts image and video quality, introducing intensity fluctuations and bright streaks influenced by lighting, camera settings, and rain properties [2]. These artifacts degrade the performance of vision tasks such as object detection [13], image recognition [15], and depth estimation, critical for applications like autonomous driving and surveillance. Rain streak removal is inherently challenging, modeled as an ill-posed problem:

$$I = C + S \tag{1}$$

where I, C, and S represent the rainy image, clean image, and rain streak, respectively. In the deraining task, the objective is to determine the clean image C given the rainy observation I. However, as there is only one equation and there are two unknowns (C and S), this is an ill-posed problem, with infinite number of solutions (in the absence of any regularization).

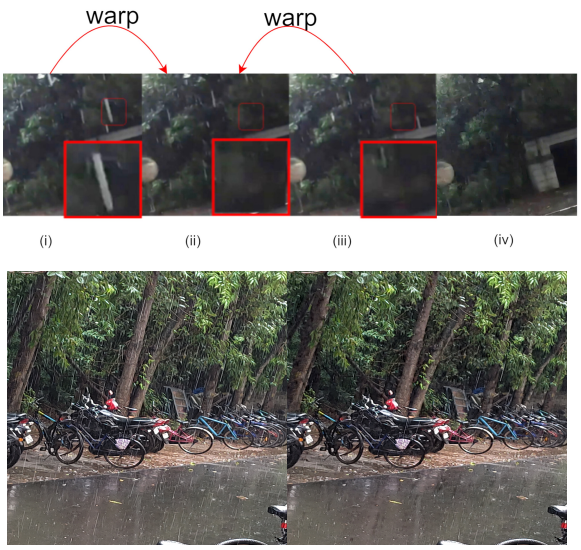


Figure 1: First row (Training): Three consecutive frames (i to iii) from a monocular video show that rain streaks seldom appear at the same scene point in adjacent frames. This enables the extraction of clean background (iv) using warping. Second row (Testing): Derained output (right) from our self-supervised method on an unseen real rainy input (left).

Traditional approaches, including sparse coding and dictionary learning [10, 12], as well as modern

deep learning methods [1, 7], have been proposed to tackle this problem. However, most methods rely on paired datasets of rainy and clean images, which are difficult to acquire and often fail to generalize well due to the domain gap between synthetic and real-world rain scenarios.

2 Proposed Solution: Self-Supervised Deraining Network (S2DNet)

We propose **S2DNet**, a self-supervised framework that removes rain streaks from single images using monocular videos captured by moving cameras. Our method leverages the dynamic nature of rain and the temporal consistency of the background across adjacent frames to align and warp them, generating pseudo-ground truth for training. The key innovation lies in utilizing depth and pose information for warping, ensuring precise scene alignment while excluding rain artifacts.

2.1 Contributions

- **Novel Framework:** First self-supervised framework leveraging depth and pose for rain streak removal, eliminating the need for paired datasets.
- **Broad Applicability:** Capable of handling moving camera scenarios, making it versatile and efficient for real-world use.
- **State-of-the-Art Results:** Extensive experiments on real-world datasets demonstrate superior performance compared to existing unsupervised methods.

3 Proposed Methodology

For a given rain image I, our target is to remove rain streaks S to yield clean pixels C. To derain single images, we leverage the temporal information present in videos. Previous models have demonstrated that there is greater scene diversity in moving videos compared to static scenes. Building on this insight, we propose warping adjacent frames of a moving video to a target frame. By exploiting the fact that rain streaks occupy different locations in adjacent frames, we devise a methodology to construct a pseudo ground truth for the clean image. We propose a deep learning model that effectively utilizes this pseudo-ground truth information to enhance the learning process for deraining. By training the model with pseudo-ground truth, we aim to improve its performance in single-image deraining. Our method allows the final model to leverage both the temporal diversity of rain streaks and the static nature of the scene, leading to superior deraining results. A detailed block diagram of our framework is given in Fig. 2.

3.1 PoseNet and DepthNet

Following [4, 3] we utilize an encoder-decoder architecture for our *PoseNet* and *DepthNet* where the latter adopts the UNet architecture [14] and the former has the same architecture as *DepthNet* except the decoder has only six outputs to get the camera transformation matrix **??**. Both the networks are trained in a self-supervised manner, employing adjacent video frames from the MRV dataset with the view consistency constraint. This ensures that the warped image (synthesized from one viewpoint to match the other) closely resembles the target image. *PoseNet* computes the relative pose between the reference frame I_i and the adjacent frames I_{i-j} (where $j \in \{-1, 1\}$) in terms of the camera transformation matrix $T_{i \to i-j}$. Subsequently, the depth D_i, adjacent frames I_{i-j}, and camera transformation matrix $T_{i \to i-j}$ are fed into the warping module. This module warps the adjacent frames to align with the reference frame, producing the warped output denoted as $I_{i-j \to i}$. This output indicates that the I_{i-j} frame has been transformed to match the perspective of the reference frame I_i. To get $I_{i-j \to i}$ we require the camera transformation matrix $T_{i \to i-j}$ between two adjacent frames from *PoseNet* and the depth D_i of the reference image I_i from *DepthNet*

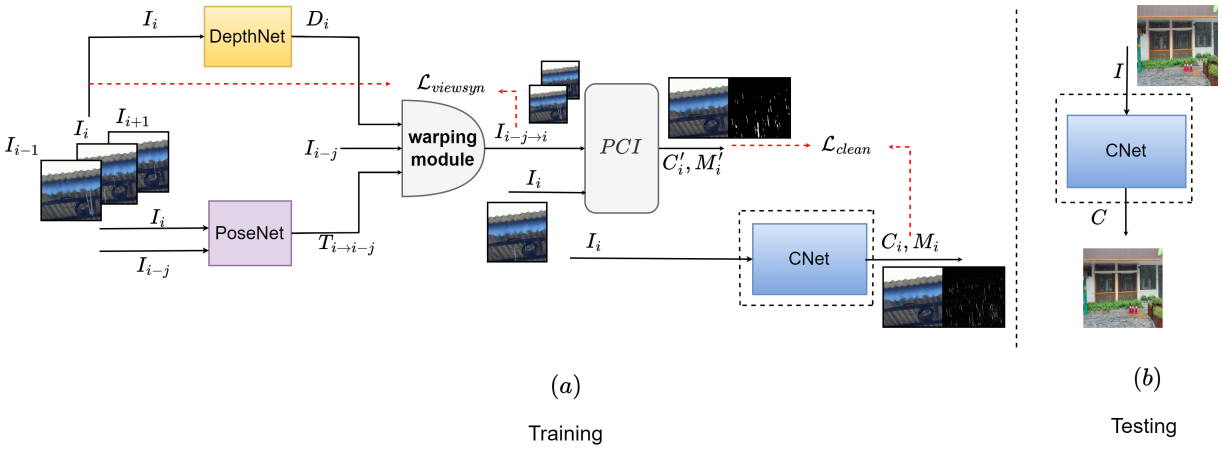


Figure 2: Complete schematic of S2DNet. DepthNet and PoseNet are trained using view-synthesis constraint. This process yields accurate depth and pose information between frames, which are then passed to the warping module. Additionally, the output of the Pseudo Clean Image (PCI) generator module serves as a surrogate for ground truth. The pseudo-ground truth aids in training CNet, which is responsible for predicting the clean images. Here $j \in \{-1, 1\}$

as given below:

$$T_{i\to i-j} = \ PoseNet(I_i, I_{i-j}) \tag{2}$$

$$D_i = \ DepthNet(I_i) \tag{3}$$

Warping is performed as

$$I_{i-j\to i} = I_{i-j}\langle \text{proj}(D_1, T_{i\to i-j}, K)\rangle \tag{4}$$

where proj() is the overall transformation matrix that maps the target coordinate x_{I_i} of I_i to the source coordinate $x_{I_{i-j\to i}}$ of $I_{i-j\to i}$ and is given by

$$x_{I_{i-j\to i}} = \text{proj}(D_i, T_{i\to i-j}, K) = KT_{1\to i}D_iK^{-1}x_{I_i} \tag{5}$$

Here K is the camera intrinsic matrix, which is the same for all the frames for a monocular video and can be found by camera calibration. $\langle\rangle$ is a sampling operator adopted from [9] to sample locally differentiable source image.

$$\mathcal{L}_{viewsyn} = \min_{j\in\{-1,1\}} \frac{\alpha}{2}\cdot(1 - SSIM(I_i, I_{i-j\to i})) + (1-\alpha)\cdot L_1(I_i, I_{i-j\to i}) \tag{6}$$

Since we train *PoseNet* and *DepthNet* using photometric constraints on the view-synthesis loss (see Eqn. 6) on the sequence of rain images (I_i, I_{i+1}.....), the warped image $I_{i+1\to i}$ closely resembles I_i. We notice that for successive rain images, the rain streaks occupy different spatial locations as compared to the reference frame (I_i). This is due to the dynamic nature of rain. It implies that $I_{i+1\to i}$ and I_i will be nearly identical except at rain streak locations. Therefore, by comparing $I_{i+1\to i}$ and I_i, we can approximate the clean pseudo background image (C_i') and binary mask M_i' for the frame I_i.

3.2 Deraining Architecture:

CNet is our sub-network designed to remove rain streaks by leveraging the spatial dependencies and directional information inherent in rain patterns. To achieve this, motivated by [8, 16], we integrate a direction-aware spatial attention block as in [8] to a residual connection block inspired by [6], ensuring that the network effectively focuses on and eliminates rain streaks while preserving the clean image. We use the pseudo-ground truth C_i' and the rain mask M_i' to train CNet. We constrain the

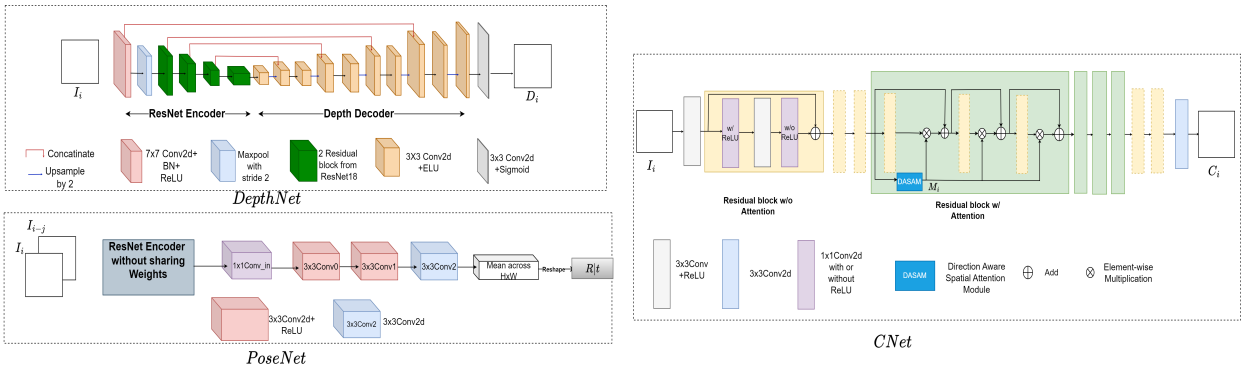


Figure 3: Architecture details of S2DNet which contains *DepthNet*, *PoseNet* and *CNet*. For *PoseNet* we use the same ResNet Encoder as in *DepthNet* but without weight sharing. Here $j \in \{-1, 1\}$. Zoom-in for better visualization.

network with a photometric loss between pseudo-ground truth and the predicted clean images, along with a mask loss and total variation loss. Mask loss with M_i' provides the necessary focus on the rain streak locations, helping the network learn the features of the rain streaks. The total variation (TV) loss helps to remove rain streaks by promoting smoothness in the output image, reducing high-frequency artifacts caused by the streaks, while preserving important structural details.

4 Conclusions

In this work, we addressed the challenge of single image deraining using a self-supervised learning approach, leveraging temporal and spatial information from monocular video frames. Our novel S2DNet framework generates pseudo-ground truth images by aligning and warping adjacent frames, enabling effective training without the need for paired rain-clean datasets. The proposed method demonstrates significant advancements over existing state-of-the-art techniques in deraining, as evidenced by extensive evaluations on multiple real-world rainy datasets. S2DNet uses only monocular videos to produce the derained image and depth map, making it more conducive and robust for practical scenarios. The innovative integration of depth and pose information for deraining in a self-supervised manner underscores the importance and novelty of our work, paving the way for future research in self-supervised learning and its applications in challenging environments.

Methods	SPA		Rain-DS		Real-1k	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DSC [12]	31.58	0.96	18.25	0.73	23.96	0.88
CycleGAN [22]	20.07	0.87	21.22	0.85	18.89	0.89
NLCL [19]	16.91	0.86	18.78	0.89	14.44	0.84
UDGNet [20]	32.23	0.97	21.97	0.79	26.2	0.94
Ours S2DNet	**33.50**	**0.977**	**25.44**	**0.904**	**29.53**	**0.953**

Table 1: Quantitative comparisons of PSNR and SSIM values for unsupervised and dictionary learning-based methods on real rain datasets (best results are shown in bold).

Rainy image DSC[12] CycleGAN NLCL[19] UDGnet[20] Ours Ground Truth

Qualitative results of unsupervised and traditional dictionary learning-based methods on Real1k [11] test dataset.

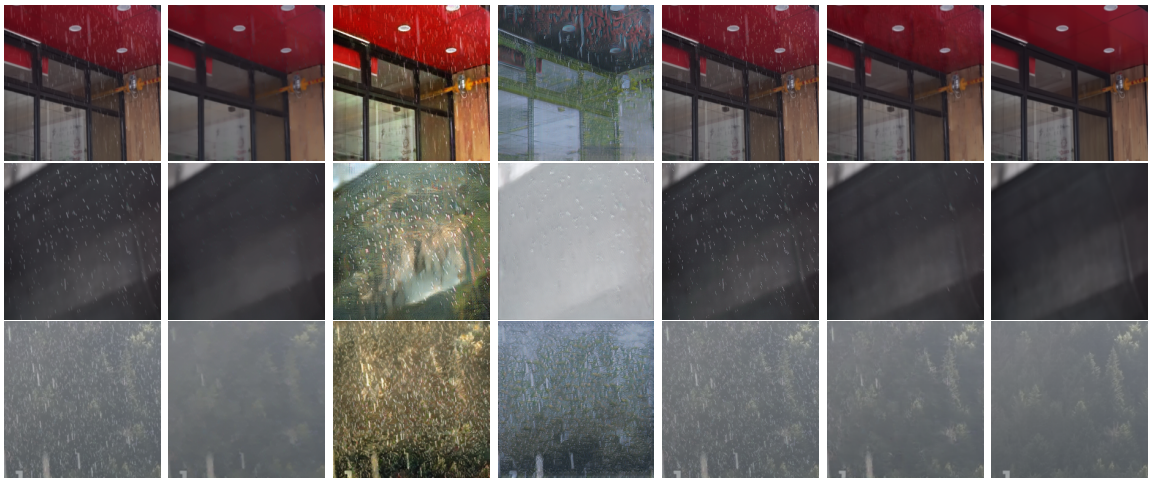
Rainy image SPANet[16] Restormer[21] Jorder[17] Syn2Real[18] Ours Ground Truth

Figure 5: Qualitative results of supervised and semi-supervised methods on LHP[5] test dataset.

Rainy image DSC[12] UDGnet[20] SPANet[16] Restormer[21] Jorder[17] Syn2Real[18]

Ours

Figure 6: Qualitative results on a real rain image randomly picked from the internet.



Rainy image DSC[12] CycleGAN NLCL[19] UDGnet[20] Ours GT

Figure 7: Additional qualitative results of unsupervised and traditional prior based methods on SPA [16] (last row) and Real-1k [11] test dataset (first two rows). Zoom-in for better visualization.

Rainy image SPANet[16] Restormer[21] Jorder[17] Syn2Real[18] Ours GT

Figure 8: Additional qualitative results with supervised models (with their pre-trained weights) on SPA[16] (first row) and Real-1k [11] test dataset (last two rows). Zoom-in for better visualization.

References

[1] Xueyang Fu et al. "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] Kshitiz Garg and Shree K Nayar. "When does a camera see rain?" In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 2. IEEE. 2005, pp. 1067–1074.

[3] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. "Unsupervised monocular depth estimation with left-right consistency". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 270–279.

[4] Clément Godard et al. "Digging into self-supervised monocular depth estimation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3828–3838.

[5] Yun Guo et al. "From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 12097–12107.

[6] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[7] Xiaowei Hu et al. "Depth-attentional features for single-image rain removal". In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2019, pp. 8022–8031.

[8] Xiaowei Hu et al. "Direction-aware spatial context features for shadow detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7454–7462.

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. "Spatial transformer networks". In: *Advances in neural information processing systems* 28 (2015).

[10] Li-Wei Kang, Chia-Wen Lin, and Yu-Hsiang Fu. "Automatic single-image-based rain streaks removal via image decomposition". In: *IEEE transactions on image processing* 21.4 (2011), pp. 1742–1755.

[11] Wei Li et al. "Toward Real-world Single Image Deraining: A New Benchmark and Beyond". In: *arXiv preprint arXiv:2206.05514* (2022).

[12] Yu Luo, Yong Xu, and Hui Ji. "Removing rain from a single image via discriminative sparse coding". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3397–3405.

[13] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany,*

October 5-9, 2015, proceedings, part III 18. Springer. 2015, pp. 234–241.

[15] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[16] Tianyu Wang et al. "Spatial attentive single-image deraining with a high quality real rain dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12270–12279.

[17] Wenhan Yang et al. "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.

[18] Rajeev Yasarla, Vishwanath A Sindagi, and Vishal M Patel. "Syn2real transfer learning for image deraining using gaussian processes". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2726–2736.

[19] Yuntong Ye et al. "Unsupervised deraining: Where contrastive learning meets self-similarity". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5821–5830.

[20] Changfeng Yu et al. "Unsupervised image deraining: Optimization model driven deep cnn". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 2634–2642.

[21] Syed Waqas Zamir et al. "Restormer: Efficient transformer for high-resolution image restoration". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 5728–5739.

[22] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. 2017.