Unsupervised Video Deraining with An Event Camera

Jin Wang Wenming Weng Yueyi Zhang* Zhiwei Xiong
University of Science and Technology of China
{jin01wang, wmweng}@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn
https://github.com/booker-max/Unsupervised-Deraining-with-Event-Camera

Abstract

Current unsupervised video deraining methods are inefficient in modeling the intricate spatio-temporal properties of rain, which leads to unsatisfactory results. In this paper, we propose a novel approach by integrating a bio-inspired event camera into the unsupervised video deraining pipeline, which enables us to capture high temporal resolution information and model complex rain characteristics. Specifically, we first design an end-to-end learning-based network consisting of two modules, the asymmetric separation module and the cross-modal fusion module. The two modules are responsible for segregating the features of the rain-background layer, and for positive enhancement and negative suppression from a cross-modal perspective, respectively. Second, to regularize the network training, we elaborately design a cross-modal contrastive learning method that leverages the complementary information from event cameras, exploring the mutual exclusion and similarity of rain-background layers in different domains. This encourages the deraining network to focus on the distinctive characteristics of each layer and learn a more discriminative representation. Moreover, we construct the first real-world dataset comprising rainy videos and events using a hybrid imaging system. Extensive experiments demonstrate the superior performance of our method on both synthetic and real-world datasets.

1. Introduction

Rain is the most common bad weather which introduces the serious degradation in captured videos and images. It not only causes the poor visual quality but also seriously deteriorates the performance of some outdoor vision tasks that assume clean video as input, *e.g.*, object tracking [15], object detection [12], person re-identification (Re-ID) [43] and segmentation [26]. Thus, it is of great importance to develop an effective video rain removal algorithm to restore the high-quality rain-free videos. Recently, many methods [42, 36, 35, 41] are proposed for rain removal and achieved significant successes in synthetic datasets. Unfortunately, most of these methods are supervised, which heavily rely on paired rain-clean data. The large domain gap between the synthetic and real rain makes them perform poorly in real-world rainy scenes.

To address this issue, the semi-supervised deraining methods [40, 8] are proposed. They commonly employed

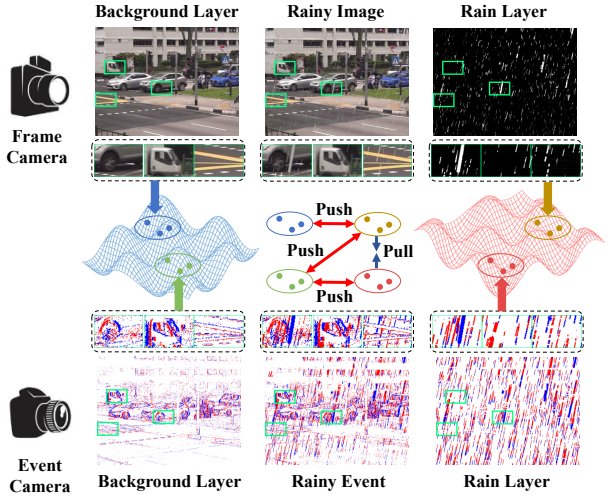


Figure 1: Our proposed cross-modal contrastive learning method includes intra-modal and inter-modal contrastive learning. In intra-modal contrastive learning, we aim to establish the mutually exclusive relationship between the rain and background layers by **pushing** them far away in both the event and frame domains. In inter-modal contrastive learning, we **pull** together the rain layer shared in two domains. Moreover, we **push** the rain layer in the frame domain and background layer in the event domain for suppressing the negative information such as the moving edges as shown in Fig. 3(f).

*Corresponding author

the labeled synthetic data for good initialization and introduce the real rains for generalization. Although the characteristics of real rains are taken into account, they cannot achieve satisfying results when the gap between synthetic and real rainy images is large. To further improve robustness, the unsupervised deraining methods have attracted more attention. Existing unsupervised deraining methods demonstrated that satisfying deraining results can be achieved by using either the temporal correlation and consistency [37] or the unpaired adversarial learning and cycle-consistency [29, 45, 39]. More recently, some methods [38, 2] exploited the underlying mutually exclusive relationship and correlation from rainy inputs to remove rains in a contrastive learning manner. Despite remarkable improvement, these frame-based methods are limited to the imaging mechanism of the conventional RGB cameras, which fail to model the complex spatio-temporal distribution of rain and present unsatisfying results.

In this paper, we introduce a novel neuromorphic sensor called event camera [4] to approach the unsupervised video deraining task. In contrast to conventional frame-based cameras that capture images at a fixed frame rate, event cameras asynchronously respond to intensity changes of each pixel in high temporal resolution and have been used for many applications [9, 30, 10, 24, 21]. We delve into the exploration of the role of event cameras in video deraining and demonstrate that event cameras can contribute to video deraining from two perspectives.

Firstly, event cameras are well-suited for modeling the complex spatio-temporal properties of rain, making it easier to distinguish between the rain layer and the background layer. The moving rain streaks produce noticeable intensity changes that match the dynamic perception of event cameras. With their high temporal resolution and high dynamic range perception of rain, event cameras can capture the fast motions of both rain and background, preserving the details of rain-free regions. Secondly, event cameras can provide complementary modality information. We can obtain both absolute intensity information and the intensity changes produced by the motion of rain and moving background objects. This way, the contrastive learning can be enhanced by exploring multiple relationships in two modality. In comparison to the conventional contrastive learning of single modality, we propose a new contrastive learning framework called cross-modal contrastive learning. It forms positive and negative pairs from both frame and event for utilizing the mutually exclusive and similar relationships between rain and background in the frame and event domains. Fig. 1 illustrates the main idea of proposed cross-modal contrastive learning. Furthermore, to enable real-world evaluations, we build a hybrid imaging system to collect a dataset of rainy videos and event streams. The main contributions are summarized as follows:

- We make the first attempt to approach unsupervised video deraining with an event camera by exploiting its effective perception of motion information.
- We formulate a cross-modal contrastive learning framework to distinguish the rain layer and background layer by exploiting their mutually exclusive and similar relationship in frame and event domains.
- We collect a real-world dataset containing rainy videos and events using a hybrid camera system.
- We achieve superior performance over existing state-of-the-art methods on both synthetic and self-collected real-world datasets.

2. Related Work

Unsupervised Deraining. Compared with the supervised deraining methods which heavily rely on labeled data and suffer from the domain gap between the synthetic and real data [3, 42, 41], unsupervised deraining methods [2, 38, 39, 29, 37] have been proposed to reduce the domain gap and improve generalization ability in real rainy scenes. Despite the corresponding merits, these frame-based methods suffer from the compromise that the imaging quality of conventional frame cameras are limited due to the low temporal resolution and low dynamic range, leading to the unsatisfying deraining results. In this work, we bring in an event camera and propose a novel cross-modal contrastive learning framework to address unsupervised video deraining.

Contrastive Learning. Contrastive learning has experienced significant progress in unsupervised representation learning. The main idea of contrastive learning is to pull close the positive pairs and push apart the negative pairs, which has been used for many applications such as image dehazing [32], image super-resolution [28] and image translation [6]. Most recently, contrastive learning has been exploited for deraining by exploring the mutual relationship between rain and background domain [2, 38]. However, the complex overlapping of the rain layer and the background layer in a single modality will make the contrastive learning difficult to converge. In contrast, we develop cross-modal contrastive learning to fully excavate the contrastive relationship from frame and event data.

Event-based Vision. Event camera has been widely used in many fields due to their unique properties of high temporal resolution, high dynamic range, and low power consumption. Recent works relevant to our paper are event-based video deblurring [31, 23], video super-resolution [13, 7] and video interpolation [33, 27]. These works introduce an event camera as an additional sensor that provides complementary information, achieving significant progress. In this paper, we make the first attempt to investigate the role of event cameras in unsupervised video deraining.

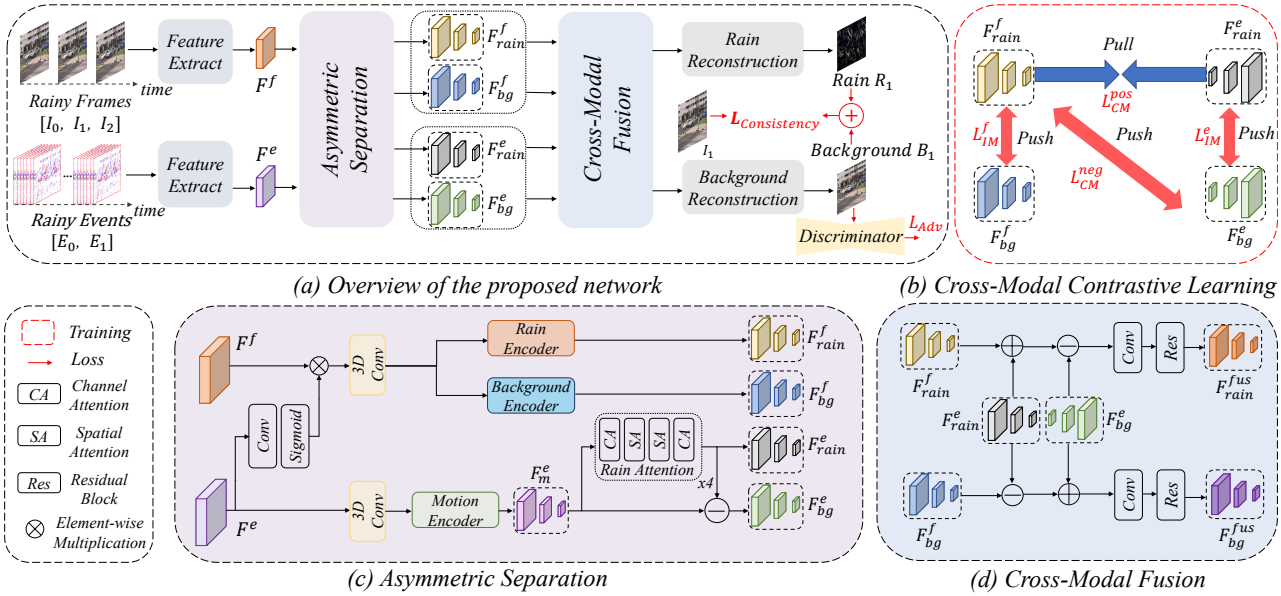


Figure 2: Pipeline of our proposed method for unsupervised video deraining with an event camera. Our method includes an event-based video deraining network and a cross-modal contrastive learning framework for regularizing the network training.

3. Deep Event-Based Video Deraining

The key of the video deraining methods is to project the rain layer and background layer into distinguishable subspaces. Existing methods struggle to achieve it by either modeling the intrinsic properties of two layers or exploiting the relationship between them. In contrast, we introduce a novel sensor, called event camera, to approach unsupervised video deraining. With the exclusive merits of high temporal resolution and high dynamic range, event cameras are capable of formulating the complex characteristics of the rain layer and offer extra supervision guidance when contrastive learning is used. Fig. 2(a) illustrates the overall pipeline of the designed event-based video deraining network, which consists of two main components: 1) asymmetric separation for segregating the features of the rain-background layer both in the event domain and frame domain, 2) cross-modal fusion for positive enhancement and negative suppression from the cross-modal perspective.

Asymmetric Separation. As shown in Fig. 2(a), we consider three consecutive frames and in-between events as the input. We convert an event sequence into a fixed-size representation $E \in \mathbb{R}^{B\times H\times W}$ according to the event representation [44]. Before the feature separation, we first extract the shallow features from three frames I_0, I_1, I_2 and two event voxel grids E_0, E_1 by two similar yet independent feature extractors, forming the frame feature F^f and event feature F^e. Note that these two features are mixed with rain and background layer features, which need to be separated.

Benefiting from the accurate motion perception of event cameras due to high temporal resolution and high dynamic range, the rain-background separation will be easier in the event domain. In contrast, the frame camera measures the absolute light intensity and texture cues. The separation of the rain streaks is more challenging in the frame domain because the rain layer exhibits strong variations (*e.g.*, direction, scale, and thickness) which always present a similar appearance to the texture of background objects. The difference between frame cameras and event cameras motivates us to adopt an asymmetric way of separating the rain layer and the background layer instead of a symmetric way with the identical feature extractor.

We illustrate the details of the asymmetric separation module in Fig. 2(c). As can be seen, the event feature F^e is first utilized to enhance the motion information in the frame feature F^f via a convolution layer followed by a sigmoid function. Then two 3D convolution layers are adopted to aggregate the temporal information for both frame and event features, which favors the latter separation. In the frame domain, we adopt two independent yet identical encoders [42] to generate multi-scale features, forming background features F^f_{bg} and rain features F^f_{rain}. In the event domain, we first use the encoder [42] to generate multi-scale motion features F^e_m, which are then fed to the rain attention block for providing the rain features F^e_{rain}. As shown in Fig. 2(c), the rain attention block is composed of repeated symmetric channel-spatial-spatial-channel attentions across the channel dimension and spatial dimension, which is able to suppress mixed information and encourages useful information relevant to rain. The features of the background layer F^e_{bg} can be obtained by subtracting F^e_{rain} from the motion features F^e_m. In such a way, we achieve rain-background separation in both frame and event domains.

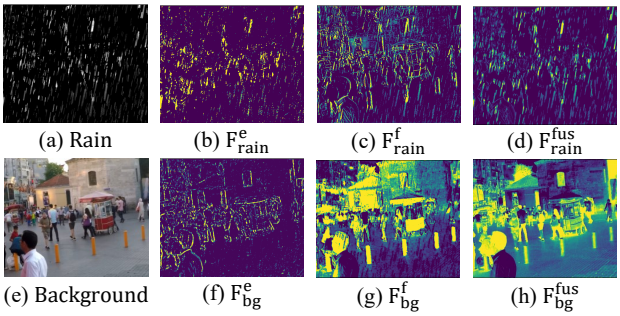


Figure 3: (a, e) The estimated rain layer R and background layer B; (b, c, f, g) four features outputted by asymmetric separation block including the rain event features, the rain frame features, the background event features, and the background frame features; (d, h) refined rain features and background features outputted by the cross-modal fusion block. Zoom in for better visualization.

Cross-Modal Fusion. Due to the exclusive advantages of event cameras and frame cameras, the complementary information can be excavated from event and frame data. As shown in Fig. 3(c), the rain feature in frame domain F^{f}_{rain} is expected to be rain information only. However, due to the limited imaging ability of frame cameras, there exists extra background noise, especially the moving edges (*e.g.*, object contour and texture boundaries) and insufficient rain information in F^{f}_{rain}. In contrast, event cameras feature high temporal resolution and high dynamic range, which leads to satisfying motion perception that favors rain-background separation. As can be seen from Fig. 3(b), (f), the rain feature in the event domain F^{e}_{rain} presents the most relevant regions about rain, and background feature F^{e}_{bg} reports accurate background information. These complementary merits of two cameras can be jointly exploited to give a clean rain feature as in Fig. 3(d).

In order to give the final outcomes of rain-background separation, we design a cross-modal fusion branch for adaptively fusing polarities (*i.e.,* positives, negatives) and cross-modal (*i.e.,* event, frame) information. As shown in Fig. 2(d), to generate the clean rain feature, we add F^{f}_{rain} and F^{e}_{rain} to enhance the rain feature, which then subtracts F^{e}_{bg} to suppress the background noise. A similar process is conducted to generate the clean background feature, in which F^{f}_{bg} is added with F^{e}_{bg} to enhance the background feature and then subtracts F^{e}_{rain} to suppress the rain noise. Mathematically, we formulate the cross-modal fusion process as:

$$\begin{aligned} F^{fus}_{rain} &= Res(Conv(F^{f}_{rain} + F^{e}_{rain} - F^{e}_{bg})), \\ F^{fus}_{bg} &= Res(Conv(F^{f}_{bg} + F^{e}_{bg} - F^{e}_{rain})). \end{aligned} \tag{1}$$

After obtaining the fused rain feature F^{fus}_{rain} and background feature F^{fus}_{bg}, we adopt two independent yet identical reconstruction blocks [42] to reconstruct the final rain layer R_1 and background layer B_1.

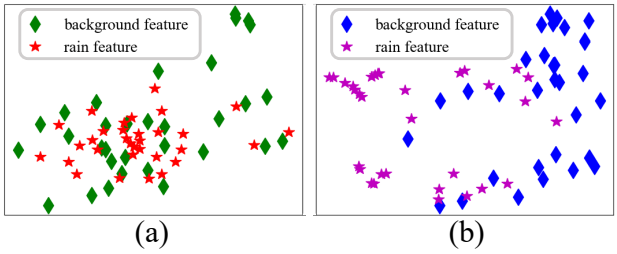


Figure 4: The t-SNE visualizations of final rain features and background features without/with the cross-modal contrastive learning in (a)/(b). The cross-modal contrastive learning is able to effectively decompose the background and rain.

4. Cross-Modal Contrastive Learning

With the proposed fully convolutional network for event-based video deraining as aforementioned in Sec. 3, we then explore how to achieve a unsupervised training scheme by utilizing the both event and frame data instead of the single modality of frame commonly-used in the prior methods [2, 29, 38].

4.1. Intra-Modal Contrastive Learning

Typically, clean background images and rainy images are significantly different. Rain streaks are characterized by sparseness, directionality and monochromaticity, while clean background images are more complex with various structures such as color, shape and texture. Therefore, whether in frame domain or in event domain, this structural discrepancy between clean background and rain streaks provides a natural guidance for separation process.

Based on these analyses, we propose to apply the intra-modal contrastive learning, which is conducted in the same modality from frame or event data to push apart rain-background features. Mathematically, given event features F^{e}_{rain}, F^{e}_{bg} and frame features F^{f}_{rain}, F^{f}_{bg}, we construct the intra-modal contrastive loss, which is formulated as:

$$\begin{aligned} \mathcal{L}^{f}_{IM} &= -\log(1 - \mathrm{sim}(F^{f}_{rain}, F^{f}_{bg})), \\ \mathcal{L}^{e}_{IM} &= -\log(1 - \mathrm{sim}(F^{e}_{rain}, F^{e}_{bg})), \\ \mathcal{L}_{IM} &= \mathcal{L}^{f}_{IM} + \mathcal{L}^{e}_{IM}, \end{aligned} \tag{2}$$

where $\mathrm{sim}(\cdot)$ denotes the operation of computing cosine similarity of two specified features. As can be seen, this intra-modal contrastive loss enforces rain-background separation in the same modality (*i.e.*, event or frame), leading to an initial outcome of deraining.

4.2. Inter-Modal Contrastive Learning

As aforementioned, the contrastive learning in the same modality provides an initial outcome of deraining. However, the same modality measurement signal shows similar properties such as structure and appearance, which confines the conditions of rain-background separation to only geometry properties such as appearance. Therefore, one problem will inevitably stands out that the background objects similar to the rain streaks would be wrongly identified as rain streaks, and be removed by mistake. It reveals that the contrastive learning in the single modality can not separate two layers effectively. To solve this challenge, we propose an inter-modal contrastive learning as a complement. In other words, the rain layer in the frame domain should not only be pushed apart from the background layer in the frame domain, but also the background layer in the event domain. In such a way, given the new observation in another modality that the background layer in event domain presents physical motion boundaries of the background objects, the non-rain region can be easily detected avoiding the misunderstanding of the background objects similar to rain streaks.

Specifically, as shown in Fig. 2(b), we construct positive pairs using the rain feature in the frame domain and the event domain. For negative pairs, we make a choice to select them from the rain feature in the frame domain and the background feature in the event domain. In addition, inspired by [34], we design a multi-scale weight-ranking way to assign the different weight to positive pairs with different similarity between multi-scale features. Mathematically, the weight is defined as:

$$W = \exp(-\alpha \cdot \mathrm{rank}(-\log(\mathrm{sim}(F^f_{rain}, F^e_{rain})))), \quad (3)$$

where α is a hyper-parameter, $\mathrm{rank}(\cdot)$ indicates the operation of sorting in descending order and getting the index, and $\mathrm{sim}(\cdot)$ denotes the operation of computing cosine similarity in feature space. The numerical values of the computed weight W is restricted in (0,1). In order to conduct inter-modal contrastive learning, we formulate the loss that is presented as:

$$\begin{aligned} \mathcal{L}^{pos}_{CM} &= -W \cdot \log(\mathrm{sim}(F^f_{rain}, F^e_{rain})), \\ \mathcal{L}^{neg}_{CM} &= -\log(1 - \mathrm{sim}(F^f_{rain}, F^e_{bg})), \\ \mathcal{L}_{CM} &= \mathcal{L}^{pos}_{CM} + \mathcal{L}^{neg}_{CM}. \end{aligned} \quad (4)$$

Note that both intra-modal and inter-modal losses are computed in the multi-scale features.

4.3. Training Loss

In addition to the proposed cross-modal contrastive loss, we further introduce extra constraints for training.

Self-Consistency Loss. To preserve the image content of the estimated background layer B_1, we utilize the self-consistency loss by composing the estimated two layers back to target rainy frame I_1, we formulate it as:

$$\mathcal{L}_{Consistency} = ||B_1 + R_1 - I_1||_1. \quad (5)$$

Adversarial Loss. We introduce the adversarial loss [5] on the predicted clean image for generating a more realistic image and keeping the data fidelity. The adversarial loss is defined as:

$$\mathcal{L}_{Adv} = \mathbb{E}_{B_1}\left[\log D(B_1)\right] + \mathbb{E}_{I_1}\left[\log(1 - D(G_B(I_1)))\right], \quad (6)$$

where $D(\cdot)$ is the discriminator, and $G_{\boldsymbol{B}}(\cdot)$ is the generator for the clean image.

The overall loss function is formulated as:

$$\mathcal{L}_{Overall} = \mathcal{L}_{Consistency} + \mathcal{L}_{Adv} + \mathcal{L}_{IM} + \mathcal{L}_{CM}. \quad (7)$$

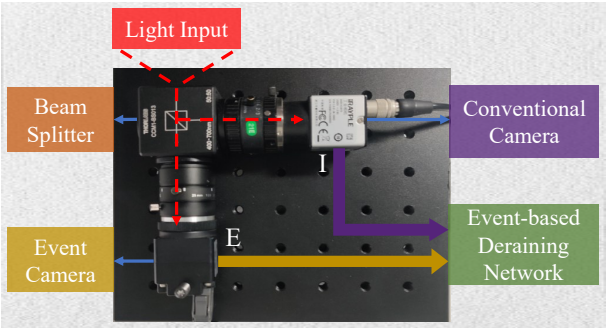


Figure 5: A hybrid camera system for building real-world rainy dataset including rainy frames and temporally synchronized event streams.

5. Experiments

5.1. Dataset Preparation

Synthetic Datasets. For training and quantitative evaluation, we generate large-scale synthetic datasets where the rain types, object motions and camera motions are considered. We choose NTURain [1], GoPro [18] and Adobe240fps [25] as the clean videos. More specially, we firstly use a video editing software to synthesize the rain layer. We randomly set the software parameters (*e.g.,* scale, thickness, wind direction, velocity, acceleration, scene depth). Then the clean videos are overlaid by the synthesized rain layers for generating rainy videos. Finally, the popular event simulator [22] is employed for generating event streams from rainy videos. In such a way, we generate four synthetic datasets: N-NTURain, N-GoProRain, N-AdobeRainH, and N-AdobeRainL, where the "N" denotes Neuromorphic, "H" and "L" denotes the dataset contains only heavy and light rain layers.

Real-World Dataset. We build a hybrid camera system to collect real-world data. As shown in Fig. 5, the hybrid camera system is composed of a conventional camera (iRAY-PLE A5031CU815 with resolution of 640×480 and a 8mm

Methods	N-NTURain		N-GoProRain		N-AdobeRainL		N-AdobeRainH	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
CUT [19]	23.50	0.8543	21.58	0.7864	24.70	0.8664	22.91	0.7698
DCD-GAN [2]	22.67	0.8065	22.02	0.7950	23.70	0.8308	21.75	0.7365
CycleGAN [45]	25.46	0.8791	25.18	0.8633	27.40	0.9125	24.45	0.8338
UDGNet [39]	29.22	0.9185	26.77	0.8824	28.97	0.9155	24.56	0.8356
NLCL [38]	28.11	0.9355	26.03	0.8921	26.99	0.9308	23.45	0.8230
DerainCycleGAN [29]	30.04	0.9242	25.07	0.8265	28.90	0.9147	22.07	0.7553
S2VD [40]	31.73	0.9347	24.88	0.8257	32.19	0.9402	27.05	0.8553
SLDNet [37]	32.74	0.9523	28.57	0.8856	32.73	0.9592	24.37	0.7727
Ours	**37.30**	**0.9756**	**32.18**	**0.9448**	**36.58**	**0.9767**	**32.35**	**0.9384**

Table 1: Quantitative comparisons on four synthetic datasets. Best in bold, the runner-up with an underline.

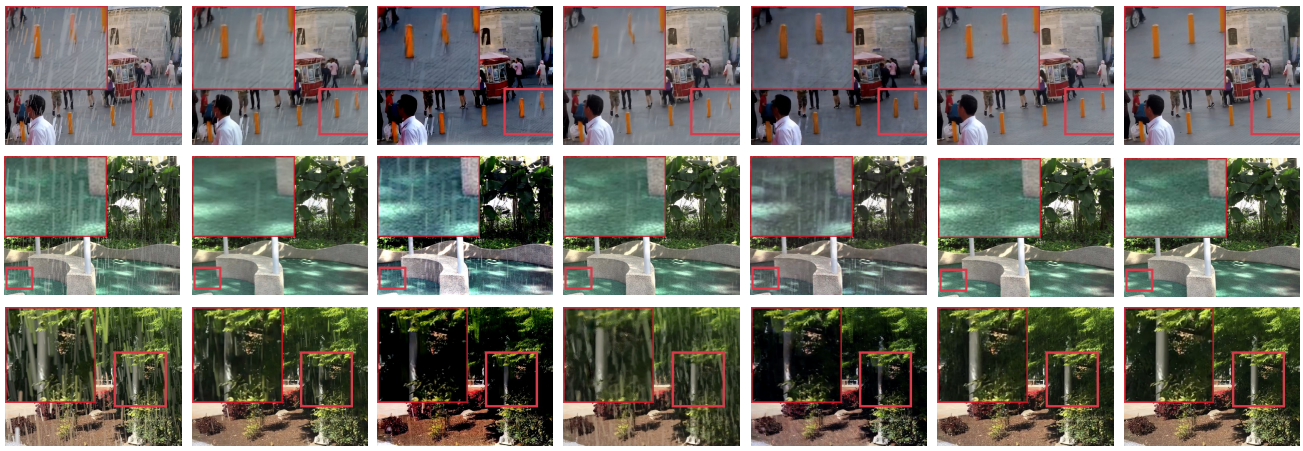


Figure 6: Qualitative comparisons of different methods on three typical frames in synthetic datasets. Frames are from N-NTURain, N-GoProRain and N-AdobeRainH, respectively. Zoom-in for better visualization.

lens), an event camera (Prophesee Gen3S1.1 with resolution of 640×480 and a 25mm lens) and a beam splitter (Thorlabs CCM1-BS013) which splits the input light and allocates them to two sensors simultaneously. For spatial-alignment, we compute the homography and radial distortion matrix between two cameras for geometric calibration. Besides, we write a synchronization script for hardware temporal synchronization. More details about the geometric calibration and temporal synchronization can be found in the Appendix.

Using the hybrid camera system, we collect a real-world dataset, called RealRain-Event. It contains 15 video sequences of varying duration from 8s to 20s in 25fps. Sequences were recorded in a variety of conditions (*e.g.*, light intensity, the field of view and camera exposure time). Besides the motion of rain streaks, the camera motion and other background objects motion have been involved in this dataset for enriching the recorded event streams. To our best, this is the **first real-world dataset** including the rainy videos and temporally synchronized event streams, which will be publicly available.

5.2. Implementation Details

We utilize the same encoder architecture as MPRNet [42] for feature extraction of both rain and background. The PatchGAN [11] is utilized for the discriminator. During the training, the original images are randomly cropped into 128×128 as input. We use the Adam optimizer [14] with the initial learning rate of 2×10^{-4}, which is steadily decreased to 0 using the cosine annealing strategy [16]. The entire network is implemented using PyTorch 1.6 [20] on two NVIDIA GTX1080Ti GPUs.

5.3. Comparisons with State-of-The-Art Methods

Baselines. We mainly select the unsupervised methods including GAN-based CycleGAN [45], contrastive-learning based CUT [19], optimization-driven UDGNet [39], self-learning video deraining method SLDNet [37], three recent unsupervised deraining methods DerainCycleGAN [29], NLCL [38] and DCD-GAN [2] and semi-supervised video deraining method S2VD [40]. We also make comparison with state-of-the-art supervised deraining method MPRNet

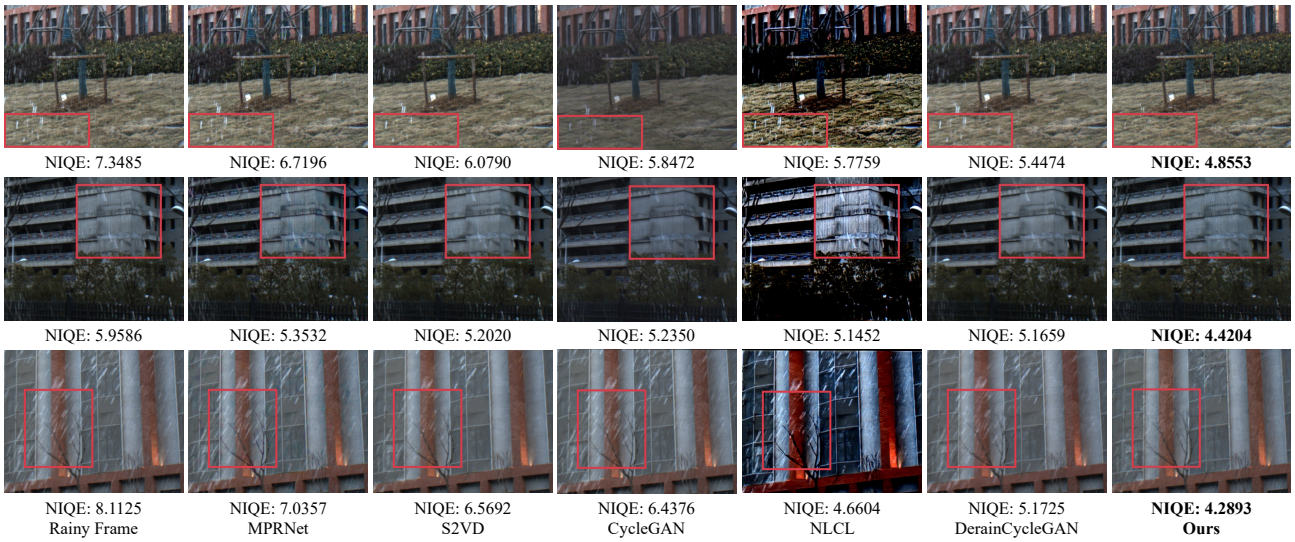


Figure 7: Qualitative comparisons on our real-world dataset RealRain-Event. Zoom-in for better visualization.

Variants	Input	Supervision	PSNR↑	SSIM↑
Model#A	Frame	Supervised	38.41	0.9820
Model#B	Frame+Event	Supervised	**41.88**	**0.9886**
Model#C	Frame	UnSupervised (w/o CMCL)	29.45	0.9207
Model#D	Frame+Event	UnSupervised (w/o CMCL)	30.00	0.9239
Model#E	Frame+Event	UnSupervised (w/ CMCL)	**37.30**	**0.9756**

Table 2: Ablation results on the effect of events. "CMCL" denotes the proposed cross-modal contrastive learning. The best results are marked in bold.

Variants	w/o $\mathcal{L}_{IM}^{f}$	w/o $\mathcal{L}_{IM}^{e}$	w/o $\mathcal{L}_{CM}^{pos}$	w/o $\mathcal{L}_{IM}^{neg}$	Full loss
PSNR↑	35.67	37.18	35.02	35.04	**37.30**
SSIM↑	0.9712	0.9752	0.9650	0.9661	**0.9756**

Table 3: Ablation results on the influence of loss functions. The best results are marked in bold.

[42] on the real rainy images. We employ the metrics of PSNR and SSIM to evaluate the deraining performance in synthetic datasets, and the non-reference natural image quality evaulator (NIQE) [17] for evaluation on the real-world dataset.

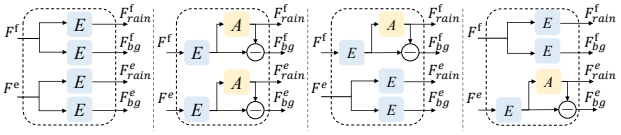
Results on Synthetic Datasets. In Tab. 1, we present the average PSNR and SSIM results on four synthetic datasets including N-NTURain, N-GoProRain, N-AdobeRainH and N-AdobeRainL. These four datasets contain various types of rains, ranging from light rain to heavy rain. The quantitative results of our proposed method achieves state-of-the-art results in terms of PSNR and SSIM, which verifies the effectiveness of our method. Especially for heavy rainy scenes, the performance of baselines are heavily limited. In addition, we provide some qualitative results of three cases in Fig. 6. The compared methods produce the incomplete

(a) Input (b) Model#C (c) Model#D (d) Model#E

Figure 8: Qualitative analysis of the effect of events. We visualize the background and rain layer generated by Model#C/Model#D/Model#E of Tab. 2 in (b)/(c)/(d).

deraining results with residual rain streaks. Besides, in the third case, DCD-GAN [2] and NLCL [38] incorrectly remove the background content due to the similarity of appearance between white pillar and rain streaks. In contrast, our method is able to effectively remove rain streaks and restore the details thanks to the accurate motion perception property of the event camera. Note that even though no extra data as inputs in the simulation process, the new generated events can be considered as motion priors of the input data, revealing the potentials of explicitly motion modeling and separation of rain and background layers. As a result, our proposed method can obtain better results on synthetic datasets both qualitatively and quantitatively.

Generalization on Real-World Dataset. For general verification in practical use, we conduct comparisons against other competitors on the RealRain-Event. For a fair comparison, we apply the pre-trained model on the N-AdobeRainH dataset of each method to remove real rain streaks. Fig. 7 presents the visual results and the corresponding non-reference metric of NIQE. It can be clearly seen that our method achieves the best visual results com-



(a) Sym: EE + EE (b) Sym: EA + EA (c) Asy: EA + EE (d) Asy: EE + EA

Figure 9: Illustration of four settings of separating the features.

Metrics	Symmetric		Asymmetric	
	EE + EE	EA + EA	EA + EE	EE + EA
PSNR↑	35.32	35.68	35.01	**37.30**
SSIM↑	0.9676	0.9638	0.9618	**0.9756**

Table 4: Ablation results on different settings of separating the features. The best results are marked in bold.

pared with other methods. We observe that the state-of-the-art supervised deraining method MPRNet [42] generates the worst result due to the large domain gap between synthetic and real rains. Unsupervised deraining methods provide a better performance compared to the supervised method in real rainy scenes. However, they fail to remove full rain streaks and preserve the details. Meanwhile, NLCL [38] brings the color distortion. In contrast, our method generates more natural and better visual deraining results, which not only removes the rain streaks but also restore the details of the background image.

5.4. Methodology Analysis

To find out what contributes to the superior performance of our approach, we conducted ablation study to demonstrate the effectiveness of each component. All ablation experiments are conducted on the N-NTURain dataset.

The Effect of Events. To validate the effectiveness of events, we choose our network as backbone and carry out the following experiments by controlling inputs, training manner and event-relevant design (*e.g.*, the proposed cross-modal contrastive learning). The quantitative results are shown in Tab. 2 and some conclusions can be made. First, for supervised manner, events bring an at least 3 dB PSNR boost by comparing Model#A and Model#B, validating the effectiveness of events. However, when adopting the unsupervised manner, events bring a slight PSNR increase by comparing Model#C and Model#D without using cross-modal contrastive learning. It reveals that simple use of events in input can not guarantee satisfying performance for unsupervised training. When cross-modal contrastive learning is exploited, an impressive PSNR gain of over 7 dB is achieved as can be seen from Model#D and Model#E. We also provide qualitative results in Fig. 8. Through these results, we draw that events indeed help video deraining but needs event-relevant designs for unsupervised training.

Variants	Cross-Modal Fusion	PSNR↑	SSIM↑
Model#A	None	34.32	0.9577
Model#B	$F^f_{rain} + F^e_{rain}$	36.22	0.9715
Model#C	$F^f_{rain} - F^e_{bg}$	36.23	0.9707
Model#D	$F^f_{rain} + F^e_{rain} - F^e_{bg} - F^f_{bg}$	35.40	0.9696
Model#E	$F^f_{rain} + F^e_{rain} - F^e_{bg}$	**37.30**	**0.9756**

Table 5: Ablation results on the strategies of cross-modal fusion. The best results are marked in bold.

The Effectiveness of the Proposed Losses. In Tab. 3, we show how each loss contributes to the final result. The $\mathcal{L}^f_{IM}$ and $\mathcal{L}^e_{IM}$ aim to decouple rain layer and background layer in event and frame modalities. The $\mathcal{L}^{pos}_{CM}$ is utilized to explore the correlations between the rain layer in frame domain and in event domain, which enables the accurate estimation of rain. The $\mathcal{L}^{neg}_{CM}$ is adopted to suppress the negative information, *i.e.*, the edge information of background objects in the separated features of rain layers. It can be observed that each loss term contributes to deraining task and the best performance is achieved by using all.

The Choice of Feature Separation. The frame camera and event camera measures rain streaks in different ways, which inspires us to employ an asymmetric feature separation network. In Tab. 4, we test different structures as shown in Fig. 9. We name "EE" as the combination of two independent encoders and "EA" as the combination of one encoder and one attention module. We use the pattern"A+B" to represent the structure of feature separation, where "A"/"B" indicates the way of feature separation in the frame/event domain. We conclude that the "EE+EA" attains the best performance, validating its effectiveness.

The Choice of Cross-Modal Fusion. Contrastive learning indicates that the positive pairs would be pulled closer and the negative pairs would be pushed away. To further enable the propagation of effective information and suppress negative information, we propose a cross-modal fusion module. In order to validate its effectiveness, we generate four variants to fuse different features in different ways. The numerical results in Tab. 2 verify that the variant of Model#E is qualified for positive enhancement and negative suppression, achieving the best result.

6. Limitation and Conclusion

Limitation. The rain streak is generally produced by the motion blur of the rain drop. The proposed method does not consider the formulation of motion blur of rain streak, which may lose efficacy when facing real complex rain scenes due to large amount of motion blur. Therefore, a new perspective of deraining via deblurring that is assisted by event cameras will be expected to solve this challenging scenes, which will be our future work.

Conclusion. This paper proposes to use event cameras for unsupervised video deraining. As validated by the comprehensive experiments, the proposed cross-modal contrastive learning boosted by event cameras demonstrates the superiority of our method on both synthetic and real-world datasets. We expect the proposed method could generalize to other bad weathers, such as snow, hail and sandstorm.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6286–6295, 2018. 5

[2] Xiang Chen, Jinshan Pan, Kui Jiang, Yufeng Li, Yufeng Huang, Caihua Kong, Longgang Dai, and Zhentao Fan. Unpaired deep image deraining using dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2017–2026, 2022. 2, 4, 6, 7

[3] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 3855–3863, 2017. 2

[4] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[6] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 746–755, 2021. 2

[7] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021. 2

[8] Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7732–7741, 2021. 1

[9] Xueyan Huang, Yueyi Zhang, and Zhiwei Xiong. High-speed structured light based 3d scanning using an event camera. *Optics Express*, 29(22):35864–35876, 2021. 2

[10] Xueyan Huang, Yueyi Zhang, and Zhiwei Xiong. Progressive spatio-temporal alignment for efficient event-based motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1537–1546, 2023. 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[12] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023. 1

[13] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Turning frequency to resolution: Video super-resolution via event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7772–7781, 2021. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023. 1

[16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *In ICLR*, 2017. 6

[17] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[18] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, July 2017. 5

[19] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 6

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[21] Yansong Peng, Yueyi Zhang, Peilin Xiao, Xiaoyan Sun, and Feng Wu. Better and faster: Adaptive event conversion for event-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2056–2064, 2023. 2

[22] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 5

[23] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video

deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2021. 2

[24] Qi Shi, Zhongfu Ye, Jin Wang, and Yueyi Zhang. Qisampling: An effective sampling strategy for event-based sign language recognition. *IEEE Signal Processing Letters*, 2023. 2

[25] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017. 5

[26] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19988–19997, 2023. 1

[27] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 2

[28] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021. 2

[29] Yanyan Wei, Zhao Zhang, Yang Wang, Mingliang Xu, Yi Yang, Shuicheng Yan, and Meng Wang. Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking. *IEEE Transactions on Image Processing*, 30:4788–4801, 2021. 2, 4, 6

[30] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 2

[31] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based blurry frame interpolation under blind exposure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1588–1598, 2023. 2

[32] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 2

[33] Zeyu Xiao, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Eva $\hat{}\{2\}$: Event-assisted video frame interpolation via cross-modal alignment and aggregation. *IEEE Transactions on Computational Imaging*, 8:1145–1158, 2022. 2

[34] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–998, 2022. 5

[35] Wending Yan, Robby T Tan, Wenhan Yang, and Dengxin Dai. Self-aligned video deraining with transmission-depth consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11966–11976, 2021. 1

[36] Wenhan Yang, Robby T Tan, Jiashi Feng, Shiqi Wang, Bin Cheng, and Jiaying Liu. Recurrent multi-frame deraining: Combining physics guidance and adversarial learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[37] Wenhan Yang, Robby T Tan, Shiqi Wang, and Jiaying Liu. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1720–1729, 2020. 2, 6

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5821–5830, 2022. 2, 4, 6, 7, 8

[39] Changfeng Yu, Yi Chang, Yi Li, Xile Zhao, and Luxin Yan. Unsupervised image deraining: Optimization model driven deep cnn. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2634–2642, 2021. 2, 6

[40] Zongsheng Yue, Jianwen Xie, Qian Zhao, and Deyu Meng. Semi-supervised video deraining with dynamical rain generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–652, 2021. 1, 6

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 2

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 1, 2, 3, 4, 6, 7, 8

[43] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023. 1

[44] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 6