# Dual-Rain: Video Rain Removal using Assertive and Gentle Teachers

Tingting Chen[1*], Beibei Lin[1⋆], Yeying Jin[1], Wending Yan[2], Wei Ye[2], Yuan Yuan[2], and Robby T. Tan[1]

[1] National University of Singapore
[2] Huawei International Pte Ltd
{tingting.c,beibei.lin,e0178303}@u.nus.edu,
{yan.wending,yewei10,yuanyuan10}@huawei.com, {robby.tan}@nus.edu.sg

**Abstract.** Existing video deraining methods addressing both rain accumulation and rain streaks rely on synthetic data for training as clear ground-truths are unavailable. Hence, they struggle to handle real-world rain videos due to domain gaps. In this paper, we present Dual-Rain, a novel video deraining method with a two-teacher process. Our novelty lies in our two-teacher framework, featuring an assertive and a gentle teacher. The novel two-teacher removes rain streaks and rain accumulation by learning from real rainy videos without the need for ground-truths. The basic idea of our assertive teacher is to rapidly accumulate knowledge from our student, accelerating deraining capabilities. The key idea of our gentle teacher is to slowly gather knowledge, preventing over-suppression of pixel intensity caused by the assertive teacher. Learning the predictions from both teachers allows the student to effectively learn from less challenging regions and gradually address more challenging regions in real-world rain videos, without requiring their corresponding ground-truths. Once high-confidence rain-free regions from our two-teacher are obtained, we augment their corresponding inputs to generate challenging inputs. Our student is then trained on these inputs to iteratively address more challenging regions. Extensive experiments show that our method achieves state-of-the-art performance on both synthetic and real-world videos quantitatively and qualitatively, outperforming existing state-of-the-art methods by 11% of PSNR on the *SynHeavyRain* dataset.

**Keywords:** Daytime Video Deraining · Assertive Teacher · Gentle Teacher

## 1 Introduction

Rain streaks and rain accumulation in rain videos can significantly degrade visibility. Rain streaks obscure background scenes, while rain accumulation washes out the scenes, resembling the effects of fog or haze. Existing video deraining methods based on deep learning [30, 38] rely on synthetic datasets for training. However, substantial domain gaps between synthetic data and real-world

[⋆] Equal Contribution

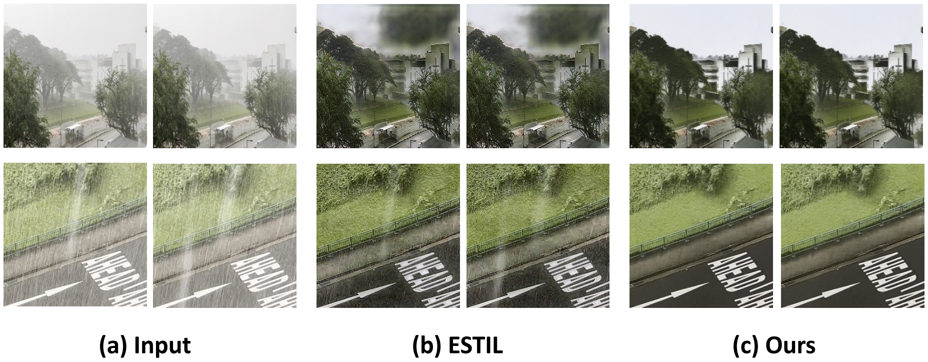


**Fig. 1:** (a): Input frames suffering from heavy rain. (b): ESTIL [38]. Third column: Our results. Our method removes both rain streaks and rain accumulation without over-suppression problem. Zoom in for better visualisation.

data impair the quality of deraining outputs. Self-learning video deraining methods [29, 33, 34] assume that rain streaks are randomly distributed across frames and obtain rain-streak-free frames by first aligning several adjacent frames. Unfortunately, this assumption does not hold for dense rain, and the method cannot handle rain accumulation.

In this paper, our main novelty is introducing a video deraining method with a novel two-teacher process that removes both rain streaks and rain accumulation in real-world videos. Our two-teacher, based on the teacher-student framework, addresses the domain gap between synthetic and real data by learning from real-world videos. Unlike the standard setup, our model employs two teachers – assertive and gentle. The assertive teacher rapidly acquires knowledge from the student through a large Exponential Moving Average (EMA) value, enhancing deraining capabilities.

This rapid learning, however, may lead to an over-suppression problem, potentially turning white objects into black, like white buildings and sky. This is because a large EMA not only accumulates the ability of deraining from the student at a fast speed, but also accumulates errors in the student quickly, resulting in inaccurate pseudo ground-truths.

To address this problem, a gentle teacher is introduced to balance the assertive teacher, preventing over-suppressed results. The gentle teacher gradually gathers knowledge from the student by employing a small EMA value, thus lessening the accumulated errors. Moreover, only high-confidence predictions generated by both teachers are utilized as the pseudo ground-truth. This allows the student to learn progressively, starting from less challenging regions and advancing to more challenging ones.

To cooperate with our novelty, the two-teacher learning process, we propose RainMix augmentation, which includes two ideas: hard and easy augmentation. Hard augmentation generates rain-specific augmentation, adding artificial rain streaks and rain accumulation into the input frames, thereby creating more challenging inputs. Note that, the rain accumulation requires depth maps for simulation. Occasional inaccuracies may arise when estimating depth maps from real-world rain videos. However, this inaccuracy can actually benefit our dual-teacher model, as it can challenge our student. Guided by pseudo ground-truths generated by our teacher, our student can learn from challenging inputs.

Easy augmentation integrates the high-confidence predictions of our two-teacher with their corresponding inputs, resulting in less challenging inputs, as they derived from easy samples (i.e., high-confidence predictions). The intuition behind our easy augmentation is that different rain effects within the same video should yield the same clear prediction. By obtaining challenging and less challenging inputs, we can train our student to progressively address more and more challenging regions. As illustrated in Fig. 1, our Dual-Rain exhibits superior performance compared to existing state-of-the-art methods. In summary, our contributions are as follows.

- We propose an assertive teacher to enable our network to address real-world rain streaks and rain accumulation. Our assertive teacher rapidly accumulates knowledge from our student through a large EMA value, thereby enhancing deraining capabilities significantly.
- We propose a gentle teacher to balance the assertive teacher, preventing over-suppressed results. Our gentle teacher gradually gathers knowledge from our student by using a small EMA value, thus lessening the accumulated errors.
- The experiments on our *SynHeavyRain* dataset demonstrate that our method outperforms existing state-of-the-art methods both quantitatively and qualitatively. Extensive experiments on real-world rain videos show that our Dual-Rain framework not only addresses rain streaks and rain accumulation effectively but also avoids the problem of over-suppression. Video deraining demos are presented in the supplementary material.

## 2 Related Works

Prior rain removal methods [4, 7, 11, 12, 19] focus on identifying the pattern of rain streaks in rainy images, which helps to differentiate the rain streak layers from the clean background layers. Existing learning-based methods use neural networks, such as the multi-stream densely-connected CNN [37], perceptual adversarial networks [26], attentive networks [27], and non-local enhanced encoder-decoder [13], to restore rainy images. Additionally, several methods address rain streaks by using priors [4, 17, 31, 35, 40].

Several existing methods for removing rain streaks and rain accumulation in single images can be applied to video rain removal [9, 31, 32, 36]. For example, Yang et al. [32] develop a CNN-based approach that addresses both rain streaks and accumulation. Hu et al. [9] remove rain effects by introducing depth-attentional features, which significantly improve the removal process. Another method by Yang et al. [31] integrates wavelet transforms within a CNN framework to eliminate rain streaks and restore background details. However, using

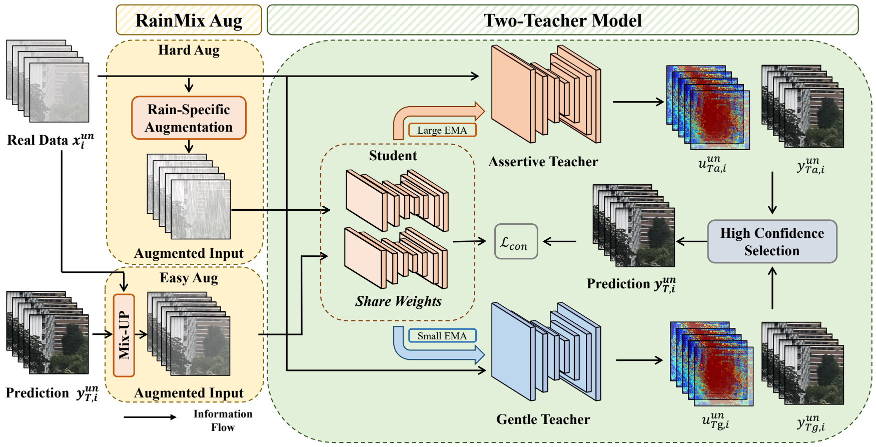


**Fig. 2:** The framework of our Dual-Rain. There are two core ideas: RainMix augmentation and two-teacher. The two-teacher consists of the assertive teacher and the gentle teacher. The assertive teacher rapidly accumulates knowledge from the student using a large EMA value, while the gentle teacher acquires knowledge slowly from the student, employing a small EMA value. The input of the student undergoes both hard augmentation and easy augmentation, generating a more challenging input and a less challenging input respectively.

single-image rain removal methods for video deraining is not effective due to the problem of flickering.

Videos contain more information than images, such as temporal correlation. Some existing video deraining methods remove rain by integrating inherent priors to distinguish between rain streaks and the clear background [1–3, 5, 10, 16, 23, 24, 28, 39]. Existing self-learning video-deraining methods [29, 33, 34] assume that rain streaks are randomly distributed across frames. These methods obtain rain-streak-free frames by initially aligning several adjacent frames with the middle frame and then integrating the aligned adjacent frames at the pixel level. However, this assumption does not hold for rain accumulation. Current self-learning methods use additional learning-based [29] or optimization-based methods [33, 34] to remove rain accumulation. Therefore, SAVD [29] is a semi-supervised deraining method, and SLDNet [33, 34] is an unsupervised deraining method.

Existing DL-based video deraining methods [6, 14, 15, 30, 38] rely on synthetic paired data for model training. Due to the domain gaps between synthetic data and real-world data, unfortunately, these methods cannot have good performance in some real-world videos. In this paper, we present a video-based framework designed to address both rain streaks and rain accumulation. We achieve this by learning from real-world videos, resulting in superior performance on data in the wild.

# 3 Proposed Method

Fig. 2 shows our pipeline, which consists of two ideas: a two-teacher and RainMix augmentation. The key novelty lies in our two-teacher process with an assertive and a gentle teacher. Our two-teacher uses real-world rainy videos to train our network, thus enabling it to address real-world rain streaks and rain accumulation effectively. The goal of RainMix augmentation is to complement our two-teacher. As shown in Fig. 2, our RainMix augmentation includes two types of augmentation: hard and easy.

Hard augmentation challenges the model with real-world rainy videos, enabling our student to gradually address more challenging regions. In contrast, easy augmentation randomly mixes our high-confidence predictions with their corresponding inputs to generate less challenging data. This generated data is then used to re-train our student, improving the deraining ability.

Initially, our two-teacher relies on a pre-trained video deraining model $w_{\text{init}}$. The main reason is that our teacher must have an initial deraining ability and thus can generate predictions with certain confidence levels. Given a synthetic dataset $\mathbf{D}_{\text{syn}} = \{\mathbf{x}_i^{\text{syn}}, \mathbf{y}_i^{\text{syn}}\}_{i=1}^{N_{\text{syn}}}$, where $\mathbf{x}_i^{\text{syn}}$ and $\mathbf{y}_i^{\text{syn}}$ are the $i$-th input and ground-truth videos, respectively. $N_{\text{syn}}$ is the number of supervised videos, we can train our network to obtain the initial network parameters $w_{\text{init}}$. These pre-trained parameters are then used to initialize the two teachers and one student.

## 3.1 Two-teacher Framework

Our two-teacher framework employs high-confidence predictions from real-world rainy videos as pseudo ground-truths to guide the student in learning the real-world deraining capability. Fig. 2 shows the whole pipeline of our two-teacher, including the assertive teacher, the gentle teacher and the student. The two teachers aim to generate predictions with confidence scores. The high-confidence predictions and their corresponding inputs are then selected to guide our student's training. Our two-teacher framework is a semi-supervised method. It is initialized with a pre-trained video deraining model that has been trained on our synthetic datasets. Then, our two-teacher framework can learn from real rainy videos without the need for ground-truths. Note that, throughout this learning process, our student model maintains access to both synthetic and real datasets at all times.

Initially, our teachers and student load the pre-trained video deraining parameters $w_{\text{init}}$. Once the models of the assertive teacher $w_{T_a}$, the gentle teacher $w_{T_g}$ and the student $w_S$ are obtained, we can begin our learning process: Given an unlabeled rainy dataset $\mathbf{D}_{\text{un}} = \{\mathbf{x}_i^{\text{un}}\}_{i=1}^{N_{\text{un}}}$, where $\mathbf{x}_i^{\text{un}}$ is the $i$-th rain video and $N_{\text{un}}$ is the number of rainy videos, we independently utilize the two teachers, $w_{T_a}$ and $w_{T_g}$, to generate predictions and confidence maps: $(\mathbf{y}_{T_a,i}^{\text{un}}, \mathbf{u}_{T_a,i}^{\text{un}})$ and $(\mathbf{y}_{T_g,i}^{\text{un}}, \mathbf{u}_{T_g,i}^{\text{un}})$. The maps $\mathbf{y}_{T_a,i}^{\text{un}}$ and $\mathbf{u}_{T_a,i}^{\text{un}}$ are predictions and confidence maps from our assertive teacher. While $\mathbf{y}_{T_g,i}^{\text{un}}$ and $\mathbf{u}_{T_g,i}^{\text{un}}$ are predictions and confidence maps from our gentle teacher.

To obtain the confidence maps, we sample regions from the input video $\mathbf{x}_i^{\text{un}}$ using a sliding window. When the window's stride is smaller than its size, the sampled regions overlap. All regions are restored using our teacher models. Ideally, the same pixel in different regions should yield the same predictions. However, variations in predictions arise because the transformer tokens for each region differ, affecting how the same pixel is represented in different contexts. We then average these overlapping predictions to obtain the teacher predictions, $\mathbf{y}^{\text{un}}T_a, i$ and $\mathbf{y}^{\text{un}}T_g, i$. Moreover, we average the predictions from our gentle and assertive teachers to obtain the final prediction $\mathbf{y}_{T,i}^{\text{un}}$.

We calculate the variance of these overlapping predictions to obtain the confidence maps $\mathbf{u}_{T_a,i}^{\text{un}}$ and $\mathbf{u}_{T_g,i}^{\text{un}}$. We then use a threshold $\mathbf{V_c}$ to convert the two confidence maps to binary masks $\mathbf{m}_{T_a,i}^{\text{un,c}}$ and $\mathbf{m}_{T_g,i}^{\text{un,c}}$. The value 1 indicates high-confidence predictions, while the value 0 means the predictions are uncertain. Moreover, we calculate the differences between $\mathbf{y}_{T_a,i}^{\text{un}}$ and $\mathbf{y}_{T_g,i}^{\text{un}}$ to obtain a different map $\mathbf{d}_{T,i}^{\text{un}}$. We also convert the different map $\mathbf{d}_{T,i}^{\text{un}}$ to a binary mask $\mathbf{m}_{T,i}^{\text{un,d}}$ using a threshold $\mathbf{V_d}$. Our final mask can be represented as $\mathbf{m}_{T,i}^{\text{un}} = \frac{\mathbf{m}_{T_a,i}^{\text{un,c}} + \mathbf{m}_{T_g,i}^{\text{un,c}}}{2} \times \mathbf{m}_{T,i}^{\text{un,d}}$.

With these confidence maps, we can collect high-confidence predictions and their corresponding inputs as paired data. To train our student, we first augment the input using our RainMix augmentation, which includes both hard and easy augmentation. Our hard augmentation adds rain-specific features, such as rain streaks and rain accumulation, to the input $\mathbf{x}_i^{\text{un}}$, obtaining the augmented input $\mathbf{x}_{i,h}^{\text{un}}$. It aims to challenge our student, forcing it to gradually address harder regions. In contrast, our easy augmentation mixes the input video $\mathbf{x}_i^{\text{un}}$ with their predictions $\mathbf{y}_{T,i}^{\text{un}}$ to generate easier inputs $\mathbf{x}_{i,e}^{\text{un}}$. The purpose of this easy input is to further improve the deraining ability. Once the hard inputs $\mathbf{x}_{i,h}^{\text{un}}$ and the easy inputs $\mathbf{x}_{i,e}^{\text{un}}$ are obtained, we use them to train our student. Note that, we only calculate the loss between the output and the prediction $\mathbf{y}_{T,i}^{\text{un}}$ in regions with a value of 1 in $\mathbf{m}_{T,i}^{\text{un}}$.

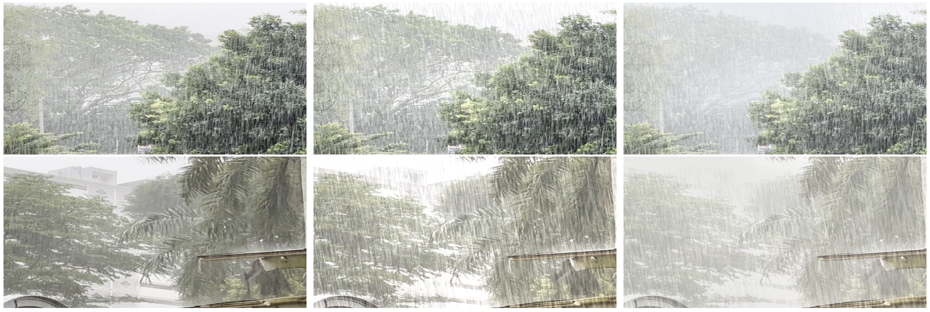
After the optimization of our student, we use the Exponential Moving Average (EMA) to update our gentle and assertive teachers. This process can be formulated as:

$$w_{T_a} = (1 - w_a^{\text{EMA}}) w_{T_a} + w_a^{\text{EMA}} w_S, \tag{1}$$

$$w_{T_g} = (1 - w_g^{\text{EMA}}) w_{T_g} + w_g^{\text{EMA}} w_S, \tag{2}$$

where $w_a^{\text{EMA}}$ and $w_g^{\text{EMA}}$ are the EMA weights of the assertive teacher and the gentle teacher, respectively. $w_a^{\text{EMA}}$ is larger than $w_g^{\text{EMA}}$.

The purposes of our assertive and gentle teachers differ. Specifically, our assertive teacher uses a large EMA value to rapidly accumulate knowledge from the student, effectively improving real-world deraining capability. However, a large EMA value also accumulates errors from the student, leading to over-suppression problems. To address this problem, we introduce a gentle teacher. It adopts a small EMA value, implying that the speed of knowledge transfer is slow. This approach tends to gather stable and accurate knowledge, which prevents



(a) Input (b) Input+S (c) Input+S+A

**Fig. 3:** Illustration of our hard augmentation. (a) shows the original real-world input frames. (b) shows the input frame with rain streaks augmentation. (c) shows the input frame with both rain streaks and rain accumulation augmentation.

the gentle teacher from overly suppressing predictions. However, relying solely on the gentle teacher may lead to sub-optimal optimization.

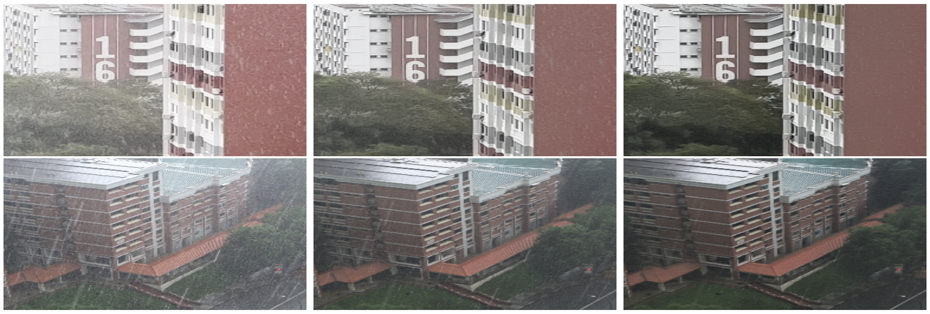
By iteratively integrating the knowledge from gentle and assertive teachers, we can generate more accurate and stable predictions that can effectively guide the student learning process. As a result, our student can pass more robust knowledge to our teacher, resulting in superior performance overall.

### 3.2 RainMix Augmentation

Our method is built on the teacher-student framework. Thus, the core idea is to use the teacher to generate a prediction $\mathbf{y}_{T,i}^{\mathrm{un}}$ from a real-world rainy video $\mathbf{x}_i^{\mathrm{un}}$. Moreover, we need to have confidence maps from the predictions and convert them into a binary mask $\mathbf{m}_{T,i}^{\mathrm{un}}$. High-confidence predictions are regarded as pseudo ground-truths to guide the learning process of the student. Specifically, we first augment the real-world rainy video $\mathbf{x}_i^{\mathrm{un}}$. We then use this augmented input along with the prediction $\mathbf{y}_{T,i}^{\mathrm{un}}$ and the mask $\mathbf{m}_{T,i}^{\mathrm{un}}$ to train our student.

The key to the student's learning process is augmentation. The main reason is that if we directly use the rainy video $\mathbf{x}_i^{\mathrm{un}}$ as input without any augmentation, the output will be the same as the prediction $\mathbf{y}_{T,i}^{\mathrm{un}}$. In this way, our student cannot learn any new knowledge from the real-world rainy video. To make the student learning process meaningful, we should apply augmentation to the input rainy video.

Motivated by this, we introduce RainMix Augmentation, a rain-specific augmentation technique. The main reason is that the augmentation should be closely aligned with our task, enabling the student to learn more effectively. To be more specific, our RainMix Augmentation includes both hard and easy augmentations. Our hard augmentation aims to challenge the rain video $\mathbf{x}_i^{\mathrm{un}}$, making it more difficult to process. For this, we develop two ways of augmentation: rain streak augmentation and rain accumulation augmentation. The appearance of



(a) Input (b) $w^{\text{Easy}} = 0.5$ (c) $w^{\text{Easy}} = 0.2$

**Fig. 4:** Illustration of our easy augmentation. (a) shows the original real-world input frames. (b) shows the the mixed input with weight $w^{\text{Easy}} = 0.5$. (c) shows the the mixed input with weight $w^{\text{Easy}} = 0.2$.

the rain streak in an image can be considered as a linear combination of a rain streak layer and the background. Therefore, our rain streak augmentation can be formulated as: $\mathbf{x}_{i,rs}^{\text{un}} = \mathbf{x}_i^{\text{un}} + \mathbf{S}_i^{\text{un}}$, where $\mathbf{x}_{i,rs}^{\text{un}}$ is the augmented image and $\mathbf{S}_i^{\text{un}}$ is the rain streak layers. Note that, rain streak layers of each frame in $\mathbf{S}_i^{\text{un}}$ are independent and randomly distributed.

In the real world, for distant scenes, rain streaks are typically not visible in distant scenes. These distant rain streaks, which come in different shapes and directions, intermingle with other water particles to form rain accumulation. Considering this property, the appearance of rain (both rain streaks and rain accumulation) can be modelled as [33]:

$$\mathbf{x}_{i,h}^{\text{un}} = \alpha(x)(\mathbf{x}_i^{\text{un}} + \mathbf{S}_i^{\text{un}}) + (1 - \alpha(x))A, \tag{3}$$

where $\mathbf{x}_{i,h}^{\text{un}}$ is the output of our hard augmentation. $\alpha(x)$ is the atmosphere transmission. $A$ is the atmospheric light. Once the augmented hard video $\mathbf{x}_{i,h}^{\text{un}}$ is obtained, we can train our student. By learning from the challenging input videos, our student can gradually address more challenging regions in real rainy videos. Fig. 3 shows the results from our hard augmentation. We can observe that after adding rain streaks and rain accumulation, the quality of the inputs becomes worse.

Our easy augmentation is proposed to further improve the deraining ability. The intuition of easy augmentation is that different rain effects with a video input should result in the same clear prediction. Motivated by this, we design our easy augmentation by mixing the input and prediction. By forcing our model to generate the same clear prediction from an easier rainy video, we further enhance the deraining ability. Our easy augmentation approach is simple: given an input rainy video $\mathbf{x}_{T,i}^{\text{un}}$, we can use our two-teacher to generate predictions $\mathbf{y}_{T,i}^{\text{un}}$. Once the input rainy video $\mathbf{x}_{T,i}^{\text{un}}$ and their corresponding predictions $\mathbf{y}_{T,i}^{\text{un}}$ are obtained, we can mix the two videos in the frame level that can be formulated as:

$$\mathbf{x}^{\text{un}}_{i,e} = w^{\text{Easy}}\mathbf{x}^{\text{un}}_{i} + (1 - w^{\text{Easy}})\mathbf{y}^{\text{un}}_{T,i}, \tag{4}$$

where $\mathbf{x}^{\text{un}}_{i,e}$ is the output of our easy augmentation. $w^{\text{Easy}}$ is a uniform distribution $(0, 1)$. As shown in Fig. 4, our easy augmentation reduces the effects of rain streaks and rain accumulation. The difficulty level of $\mathbf{x}^{\text{un}}i, e$ can be adjusted by the value of $w^{\text{Easy}}$.

## 4 Experimental Results

We first discuss the implementation details of our experiments. To demonstrate the effectiveness of our novel two-teacher learning process, we then provide comparisons between our method and state-of-the-art methods using both synthetic and real-world data. Finally, we conduct ablation studies to further demonstrate the effectiveness of each component.

***Datasets*** We compare our methods with state-of-the-art methods on the synthetic datasets *SynHeavyRain* and *Haze-NTU*. Among these, *SynHeavyRain* is collected by us and comprises 11 videos for training and 9 videos for testing. Each video contains more than 100 frames. We add rain accumulation to *NTU-Rain* [6] to create *Haze-NTU* since *NTURain* only contains rain streaks. The rain accumulation synthesis method is detailed in our supplementary material. The frames and videos are (∼100, 16) for *Haze-NTU*. We will also compare the performance on real-world videos collected by ourselves.

***Implementation Details*** Our Dual-Rain includes two core ideas: a two-teacher framework and RainMix augmentation. Our pre-trained network is implemented using a transformer-based architecture [22], which includes 10 transformer blocks. Each transformer block consists of a multi-head self-attention module and an MLP layer. However, it is designed for image-based tasks. To handle video inputs, we replace the 2D patch embedding (PE) with 3D PE to extract tokens from videos, add spatiotemporal positional encoding, and use transformer layers to process these tokens. The loss for our pre-trained model is the mean absolute error without regularization. The hidden features of both the multi-head self-attention module and MLP layer are set to 768. The patch size of our video transformer is set to 4. The input and the output of our network are set to 16 frames.

Our two-teacher framework consists of an assertive teacher and a gentle teacher. Both assertive and gentle teachers are initialized by the pre-trained parameters. To obtain binary masks and different maps, we set $\mathbf{V_c}$ and $\mathbf{V_d}$ to 0.05. The EMAs of the assertive teacher and the gentle teacher are set to 0.003 and 0.001, respectively. In this case, The assertive teacher acquires knowledge 3x faster than the gentle teacher. For our RainMix augmentation, $W^{\text{Easy}}$ is a uniform distribution (0.5, 1). We use the Adam optimizer to update the network

**Table 1:** Quantitative comparison on *SynHeavyRain* and *Haze-NTU*. SLDNet [34], BIPNet [8], SAVD [29] and ESTIL [38] are CNN-based methods. WDiff [21] and V-DiT [22] are diffusion-based methods.

| Datasets | Metrics | Video Swin'22 | BIPNet'22 | ESTIL'22 | WDiff'23 | SLDNet'20 | SAVD'21 | V-DiT'23 | **Ours** |
|---|---|---|---|---|---|---|---|---|---|
| SynHeavyRain | PSNR | 19.53 | 20.51 | 20.43 | 20.57 | 13.48 | 16.55 | 21.39 | **23.89** |
| | SSIM | 0.7071 | 0.7128 | 0.7458 | 0.8009 | 0.6412 | 0.6204 | 0.6867 | **0.8567** |
| Haze-NTU | PSNR | 17.34 | 17.86 | 16.91 | 16.97 | 17.36 | 17.79 | 18.22 | **19.06** |
| | SSIM | 0.7268 | 0.7052 | 0.6468 | 0.7802 | 0.6743 | 0.7165 | 0.6409 | **0.7847** |

parameters of our student model. The learning rate is set to 1e-4. For the pre-training stage and our two-teacher stage, the training batch sizes are set to 12 and 4, respectively. All experiments are conducted on four A5000 GPUs.

### 4.1 Comparison on Synthetic Datasets

We compare our method with the following state-of-the-art-methods: Video Swin transformer [18], BIPNet [8], ESTIL [38], WDiff [21], SLDNet [34], SAVD [29], as well as V-DiT. Note that, V-DiT is a 3D version implemented by us, based on DiT [22]. To implement V-DiT, we transition the 2D transformer to a 3D variant and employ rain videos as the condition. Then, we train V-DiT using a conditional diffusion method on our datasets. Video-swin is a transformer-based method. We remove the spatial downsampling of Video-swin and train it on our datasets. We re-implement the SAVD code ourselves. SLDNet, BIPNet, ESTIL and SAVD are CNN-based methods, where SLDNet, ESTIL and SAVD are designed for video deraining. WDiff and V-DiT are diffusion-based methods. We re-train these methods on our *SynHeavyRain* dataset for a fair comparison.

***Quantitative Evaluation*** Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are used to evaluate the quality of restored videos on the synthetic datasets. The experimental results are shown in Table 1. It can be observed that our Dual-Rain method outperforms other state-of-the-art approaches. Compared with diffusion-based deraining methods, such as WDiff and V-DiT, our method outperforms V-DiT by 11% in PSNR and 24% in SSIM, and outperforms WDiff by 16% in PSNR and 6% in SSIM on the *SynHeavyRain* dataset. For existing self-learning video deraining methods, such as SLDNet and SAVD, our Dual-Rain framework outperforms SLDNet by 10.41 dB and SAVD by 7.34 dB, respectively. For methods that rely on synthetic datasets for training, such as Video-swin, BIPNet, and ESTIL, our Dual-Rain framework outperforms Video-swin by 4.36 dB, BIPNet by 3.38 dB, and ESTIL by 3.46 dB, respectively. Experiments on *Haze-NTU* also demonstrate the effectiveness of our two-teacher framework.

### 4.2 Comparison on Real-World Datasets

***Quantitative Evaluation on Non-Reference Metrics*** In this section, we evaluate the quality of real-world restored rain videos. Since we do not have the

**Table 2:** Quantitative comparisons on non-reference metrics BRISQUE and Hyper-IQA. We compare our method with following state-of-the-art methods: BIPNet [8], V-DiT [22], SAVD [29], SLDNet [33, 34], and ESTIL [38]. SAVD is a semi-supervised video deraining method, while SLDNet is an unsupervised video deraining method.

| Metrics | BIPNet | V-DiT | SAVD | SLDNet | ESTIL | **Ours** |
|---|---|---|---|---|---|---|
| BRISQUE ↓ | 32.34 | 31.06 | 29.49 | 30.35 | 28.46 | **24.32** |
| HyperIQA ↑ | 0.3430 | 0.5667 | 0.6695 | 0.6536 | 0.5833 | **0.6702** |

ground truth of real rain videos, we use two non-reference metrics, BRISQUE [20] and HyperIQA [25], to compare the performance of our methods with other state-of-the-art methods. The experimental results are shown in Table 2. We observe that the existing semi-supervised deraining method SAVD [29] and the unsupervised deraining method SLDNet [33, 34] achieve BRISQUE scores of 29.49 and 30.35, respectively. In contrast, our method achieves a BRISQUE score of 24.32, significantly enhancing the quality of restored videos.

***Dense Rain Streaks*** To verify the effectiveness of our Dual-Rain framework, we firstly evaluate our methods on real-world videos with dense rain streaks. We compare our performance with the first stage of SAVD [29] and ESTIL [38]. SAVD is a self-leaning method which removes rain streaks in videos and ESTIL is a video deraining method trained on synthetic datasets. The experimental results are shown in Fig. 5. For each method, we show the results from three consecutive frames.

As shown in Fig. 5 (b-c), It can be observed that existing methods, such as ESTIL and SAVD, cannot effectively remove dense rain streaks. There are still some rain streaks remaining in the street. This is because SAVD is a self-learning video deraining method, assuming that rain streaks are randomly distributed over frames. It attempts to recover the clean middle frame by integrating aligned adjacent frames. However, in situations with dense rain streaks, this assumption can fail. Specifically, in videos with dense rain streaks, the adjacent frames may not contain rain-free information for the current frame, leading to inaccurate rain streak removal. As shown in Fig. 5 (c), SAVD cannot effectively remove rain streaks in the street. For ESTIL, it is trained on synthetic datasets. However, there is a large domain gap between synthetic data and real-world data. Consequently, models trained on these synthetic datasets may not accurately model the distribution of clear background, leading to inaccurate rain streak removal. As shown in Fig. 5 (b), ESTIL also suffers from the inaccurate rain streak removal problem.

In contrast, our dual-teacher framework utilizes real-world rain videos to enable our model to effectively address both real-world rain streaks and rain accumulation. To be more specific, we introduce a teacher-student learning process. The teacher model generates pseudo ground truths for real-world rain videos. Once high-confidence predictions are obtained, they are used to train our student model. By iteratively learning from high-confidence predictions, our student

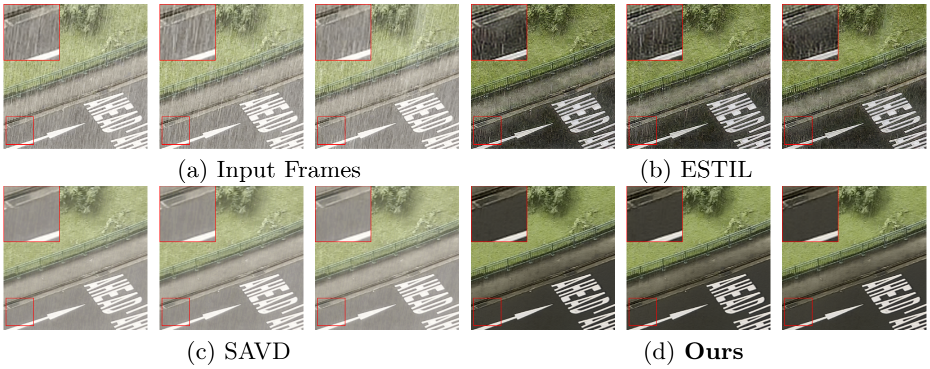
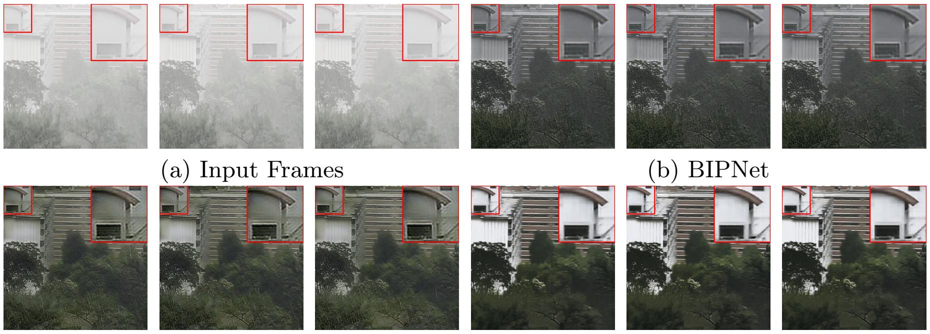


**Fig. 5:** Qualitative comparison on real-world data with dense rain streaks. SAVD [29] cannot remove rain streaks entirely due to the assumption that rain streaks are randomly distributed. ESTIL [38] fails to remove all rain streaks since it relies on synthetic data for training. In contrast, our method generates more effective results. Zoom in for better visualisation.

model can gradually address more challenging cases, such as regions suffering from dense rain streaks or accumulation. As shown in Fig. 5, our dual-teacher framework achieves significant performance improvements in real-world dense rain streak removal.

***Dense Rain Accumulation*** Compared to rain videos with dense rain streaks, those with dense rain accumulation are more complex. Therefore, we have also collected and evaluated our methods on these dense rain accumulation videos. The experimental results are shown in Fig. 6 and Fig. 7. For each method, three consecutive frames are shown. We also provide video deraining demos for all methods in our supplementary material. We compare our method with state-of-the-art methods, including BIPNet [8] and ESTIL [38]. It can be observed that existing methods fail to restore clean videos from heavy rain videos. As shown in the figures, they cannot remove some rain streaks and also suffer from a color shift problem, especially on the white building. This is because these methods rely on synthetic datasets for training. However, there is a large domain gap between synthetic datasets and real-world rain videos.

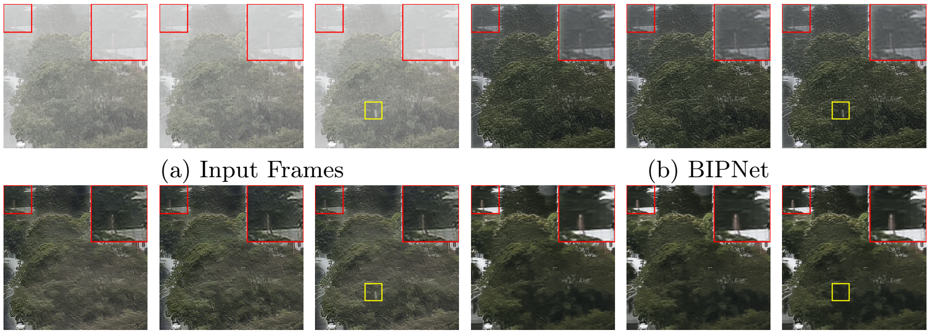
In contrast, our Dual-Rain framework allows our model to learn from real-world rain videos, thereby enhancing the deraining capabilities. Specifically, our framework includes an assertive teacher and a gentle teacher. The assertive teacher rapidly gathers knowledge from the student, significantly enhancing the model's real-world deraining capabilities. However, this strategy may also lead to the quick accumulation of student errors. To address this problem, we introduce a gentle teacher to balance the assertive teacher and prevent over-suppressed results. As shown in Fig. 6 and Fig. 7, our Dual-Rain method effectively restores the clean background of rain videos with dense rain streaks and rain accumulation. More visual results and videos are provided in the supplementary material.



(a) Input Frames (b) BIPNet

(c) ESTIL (d) **Ours**

**Fig. 6:** Qualitative comparison on real-world data with dense rain accumulation. As highlighted in red boxes, BIPNet [8] and ESTIL [38] fail to recover the color of the white building and other details. Besides, both of the methods tend to generate faded color. Zoom in for better visualisation.



(a) Input Frames (b) BIPNet

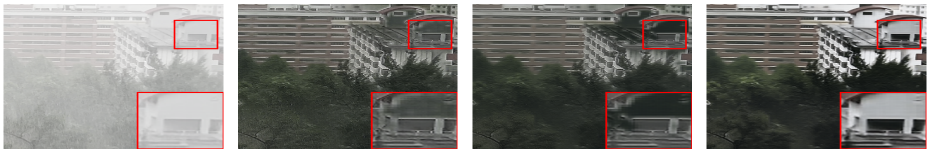
(c) ESTIL (d) **Ours**

**Fig. 7:** Qualitative comparison on real-world data with dense rain accumulation. As highlighted in yellow boxes, BIPNet [8] and ESTIL [38] fail to remove rain streaks. Besides, as highlighted in red boxes, our method also produces more clean background compared to the two methods. Zoom in for better visualisation.

### 4.3 Ablation Studies

In this section, we conduct several ablation studies on the real-world videos to evaluate the effectiveness of our two-teacher model. The experimental results are shown in Fig. 8. As indicated in Fig. 8 (b), the result from the pre-trained model does not perfectly recover the clean background, especially in the region of white building. This is because the pre-trained model is trained on the synthetic datasets, which suffers from the domain gaps between the synthetic data and real-world data, thus generating inaccurate results.

Fig. 8 (c) displays the outcome of the only assertive teacher model based on pre-training parameters. In comparison with (b), the model exhibits a more robust ability for rain removal. However, the result in the region of the white



(a) Input (b) P (c) P+AT (d) P+AT+GT

**Fig. 8:** Ablation studies of our two-teacher model. (a) shows the original real-world input frame. (b) shows the result of the pre-training model (P), which is trained on synthetic datasets. (c) shows the result of the only assertive teacher model (AT) based on pre-training parameters (P). (d) shows the result of our two-teacher model with both assertive teacher (AT) and gentle teacher (GT), based on pre-training parameters (P).

building becomes even worse. This is because the assertive teacher learns knowledge from the student model in a rapid way and also accumulates the error from the student model quickly at the same time.

Our two-teacher model, as shown in Fig. 8 (d), achieves better performance compared to the model with only an assertive teacher, especially in the white building region. This is because our gentle teacher gathers knowledge slowly from the student, resulting in an sub-optimal but stable optimization, thus preventing the results from over-suppression problems. These experiments demonstrate that each part of our two-teacher model contributes to the superior results.

## 5 Conclusion

In this paper, we propose Dual-Rain, a novel video deraining framework with a two-teacher framework. Our main novelty is our two-teacher framework, which includes an assertive teacher and a gentle teacher. The assertive teacher learns knowledge from the student model rapidly, acquiring the ability of rain removal quickly, but also suffers from an over-suppression problem. In contrast, the gentle teacher gains knowledge from the student model slowly, leading to sub-optimal but more stable results. Guided by the pseudo ground-truths obtained from our assertive and gentle teachers, our student iteratively learns from high-confidence predictions and gradually addresses more challenging real-world rain regions. In addition, the student model is challenged with augmented data using our RainMix augmentation, which incorporates both hard and easy augmentation. Hard augmentation involves rain-specific augmentation, while easy augmentation combines rainy inputs with pseudo ground-truths. Thanks to our hard and easy augmentations, our student model can effectively learn from unlabeled rain videos, enhancing the visibility of real-world rain videos. Extensive experiments demonstrate that our method achieves state-of-the-art performance quantitatively and qualitatively.

## References

1. Barnum, P.C., Narasimhan, S., Kanade, T.: Analysis of rain and snow in frequency space. International journal of computer vision **86**, 256–274 (2010) 4
2. Bossu, J., Hautiere, N., Tarel, J.P.: Rain or snow detection in image sequences through use of a histogram of orientation of streaks. International journal of computer vision **93**, 348–367 (2011) 4
3. Brewer, N., Liu, N.: Using the shape characteristics of rain to identify and remove rain from video. In: Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, SSPR & SPR 2008, Orlando, USA, December 4-6, 2008. Proceedings. pp. 451–458. Springer (2008) 4
4. Chang, Y., Yan, L., Zhong, S.: Transformed low-rank model for line pattern noise removal. In: Proceedings of the IEEE international conference on computer vision. pp. 1726–1734 (2017) 3
5. Chen, J., Chau, L.P.: A rain pixel recovery algorithm for videos with highly dynamic scenes. IEEE transactions on image processing **23**(3), 1097–1104 (2013) 4
6. Chen, J., Tan, C.H., Hou, J., Chau, L.P., Li, H.: 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. pp. 6286–6295 (2018) 4, 9
7. Chen, Y.L., Hsu, C.T.: A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In: Proceedings of the IEEE international conference on computer vision. pp. 1968–1975 (2013) 3
8. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burst image restoration and enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5759–5768 (2022) 10, 11, 12, 13
9. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8022–8031 (2019) 3
10. Jiang, T.X., Huang, T.Z., Zhao, X.L., Deng, L.J., Wang, Y.: A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 4057–4066 (2017) 4
11. Kang, L.W., Lin, C.W., Fu, Y.H.: Automatic single-image-based rain streaks removal via image decomposition. IEEE transactions on image processing **21**(4), 1742–1755 (2011) 3
12. Kim, J.H., Lee, C., Sim, J.Y., Kim, C.S.: Single-image deraining using an adaptive nonlocal means filter. In: 2013 IEEE international conference on image processing. pp. 914–917. IEEE (2013) 3
13. Li, G., He, X., Zhang, W., Chang, H., Dong, L., Lin, L.: Non-locally enhanced encoder-decoder network for single image de-raining. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1056–1064 (2018) 3
14. Liu, J., Yang, W., Yang, S., Guo, Z.: D3r-net: Dynamic routing residue recurrent network for video rain removal. IEEE Transactions on Image Processing **28**(2), 699–712 (2018) 4
15. Liu, J., Yang, W., Yang, S., Guo, Z.: Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3233–3242 (2018) 4
16. Liu, P., Xu, J., Liu, J., Tang, X.: Pixel based temporal analysis using chromatic property for removing rain from videos. Comput. Inf. Sci. **2**(1), 53–60 (2009) 4

17. Liu, R., Jiang, Z., Fan, X., Luo, Z.: Knowledge-driven deep unrolling for robust image layer separation. IEEE transactions on neural networks and learning systems **31**(5), 1653–1666 (2019) 3
18. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022) 10
19. Luo, Y., Xu, Y., Ji, H.: Removing rain from a single image via discriminative sparse coding. In: Proceedings of the IEEE international conference on computer vision. pp. 3397–3405 (2015) 3
20. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012) 11
21. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 10
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 9, 10, 11
23. Ren, W., Tian, J., Han, Z., Chan, A., Tang, Y.: Video desnowing and deraining based on matrix decomposition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4210–4219 (2017) 4
24. Santhaseelan, V., Asari, V.K.: Utilizing local phase information to remove rain from video. International Journal of Computer Vision **112**, 71–89 (2015) 4
25. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3667–3676 (2020) 11
26. Wang, C., Xu, C., Wang, C., Tao, D.: Perceptual adversarial networks for image-to-image transformation. IEEE Transactions on Image Processing **27**(8), 4066–4079 (2018) 3
27. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: 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. pp. 12270–12279 (2019) 3
28. Wei, W., Yi, L., Xie, Q., Zhao, Q., Meng, D., Xu, Z.: Should we encode rain streaks in video as deterministic or stochastic? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2516–2525 (2017) 4
29. Yan, W., Tan, R.T., Yang, W., Dai, D.: Self-aligned video deraining with transmission-depth consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11966–11976 (2021) 2, 4, 10, 11, 12
30. Yang, W., Liu, J., Feng, J.: Frame-consistent recurrent video deraining with dual-level flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 1, 4
31. Yang, W., Liu, J., Yang, S., Guo, Z.: Scale-free single image deraining via visibility-enhanced recurrent wavelet learning. IEEE Transactions on Image Processing **28**(6), 2948–2961 (2019) 3
32. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1357–1366 (2017) 3

33. Yang, W., Tan, R.T., Wang, S., Kot, A.C., Liu, J.: Learning to remove rain in video with self-supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 2, 4, 8, 11
34. Yang, W., Tan, R.T., Wang, S., Liu, J.: 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. pp. 1720–1729 (2020) 2, 4, 10, 11
35. Yasarla, R., Patel, V.M.: Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8405–8414 (2019) 3
36. Yasarla, R., Sindagi, V.A., Patel, V.M.: Syn2real transfer learning for image deraining using gaussian processes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2726–2736 (2020) 3
37. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 695–704 (2018) 3
38. Zhang, K., Li, D., Luo, W., Ren, W., Liu, W.: Enhanced spatio-temporal interaction learning for video deraining: faster and better. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1287–1293 (2022) 1, 2, 4, 10, 11, 12, 13
39. Zhang, X., Li, H., Qi, Y., Leow, W.K., Ng, T.K.: Rain removal in video by combining temporal and chromatic properties. In: 2006 IEEE international conference on multimedia and expo. pp. 461–464. IEEE (2006) 4
40. Zhu, L., Fu, C.W., Lischinski, D., Heng, P.A.: Joint bi-layer optimization for single-image rain streak removal. In: Proceedings of the IEEE international conference on computer vision. pp. 2526–2534 (2017) 3