Variational Image Deraining

Yingjun Du[1,2,*] Jun Xu[3] Qiang Qiu[4] Xiantong Zhen[2,5,+] Lei Zhang[5]
[1]University of Amsterdam, Amsterdam, Netherlands,
[2]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[3]College of Computer Science, Nankai University, Tianjin, China, [4]Duke University, Durham, USA
[5]Guangdong University of Petrochemical Technology, Guangdong, China
{duyingjun2018, nankaimathxujun, zhenxt, zhangleihrbeu}@gmail.com, qiang.qiu@duke.edu

Abstract

Images captured in severe weather such as rain and snow significantly degrade the accuracy of vision systems, e.g., for outdoor video surveillance or autonomous driving. Image deraining is a critical yet highly challenging task, due to the fact that rain density varies across spatial locations, while the distribution patterns simultaneously vary across color channels. In this paper, we propose a variational image deraining (VID) method by formulating image deraining in a conditional variational auto-encoder framework. To achieve adaptive deraining to spatial rain density, we generate a density estimation map for each color channel, which can largely avoid over and under deraining. In addition, to address cross-channel variations, we conduct channel-wise deraining, motivated by our observation that bright pixels do not tend to remain bright after deraining unless their color channels are handled separately. Experimental results show that the proposed deraining method achieves superior performance on both synthesized and real rainy images, surpassing previous state-of-the-art methods by large margins.

1. Introduction

Rain streaks on an image greatly degrade its visual quality, and produce significant obstacles to computer vision algorithms. Therefore, image deraining has recently received increasing attention due to its prerequisite role in many vision tasks, such as video surveillance [24], object detection [8], and object tracking [28] in autonomous driving.

However, image deraining is highly non-trivial for three main reasons. First, it is difficult to define the optimal solution for single image deraining due to its inherent ill-posed nature. Moreover, most existing methods simply model a deterministic function in deraining tasks, which may incorrectly collapse to a single-modal distribution. Second, rain streaks are usually not evenly distributed on a rainy image, i.e., rain density spatially varies across the rainy image. Deraining by treating spatial locations equally inevitably results in over or under deraining [6, 34, 16]. To tackle this problem, Zhang et al. [36] proposed density-aware deraining method by utilizing global rain density information via a multi-stream dense network. But this method still produces inaccurate deraining results in local regions. Third, the rain density distribution pattern usually varies dramatically across the three color channels, which is ignored by previous methods. As a result, the pixel brightness in individual channels will be largely compromised during deraining without processing the three channels separately.

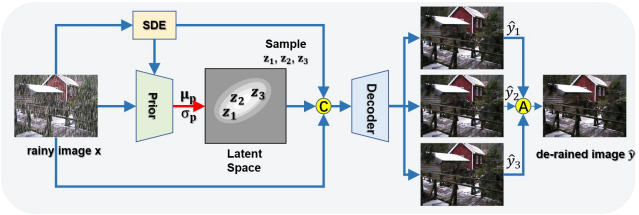


Figure 1: **Illustration of the inference stage of VID.** We sample multiple latent variables $\mathbf{z}_p$ from the prior distribution $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p)$ and adopt the Monte Carlo method [22] to perform a deterministic inference followed by weighted averaging. The red arrow indicates the process of generating latent space. ⓒ denotes the concatenation, "A" denotes the average operation. SDE means the spatial density estimation module, which will be introduced in §3.4.

In this paper, to address the aforementioned challenges, we propose a variational image deraining (VID) approach within the framework of conditional variational auto-encoder (CVAE) [23]. The CVAE provides a well-suited learning architecture for image deraining due to its strong capability to model the latent distribution of image priors, from which the corresponding clean image can be generated. Different from modeling a deterministic map-

* Work done as an intern at IIAI; + Corresponding Authors

ping function in previous methods, we propose to simultaneously learn the latent representation of a clean image and predicts multiple possible derained images through a feasible CVAE model. In the learning stage, conditioned on a rainy image, the encoder learns to map its corresponding clean image into a latent distribution that shares common information for clean images, while the decoder reconstructs the derained images based on a sampled variable from the latent space distribution. In the inference stage, we sample multiple latent variables from the prior distribution using the prior network and adopt the Monte Carlo method [22] to perform deterministic inference, as shown in Fig. 1.

Moreover, we observed that the derained images remain bright in the regions corrupted by rain streaks. The reason is that, since the rain density distribution differs across different color channels, joint processing the three channels in the same manner will be problematic. By processing each channel individually in our CVAE method, we experimentally found that this problem can be largely alleviated. The advantages of exploring different color channels have already been shown for other low-level image processing tasks such as dehazing [9] and denosing [30, 31]. The proposed channel-wise deraining strategy is able to accurately preserve the brightness of the derained images, which has been demonstrated rigorously under the bright channel prior (BCP) [7]. Note that the channel-wise processing is coupled with the proposed CVAE model, thus enabling it to learn a unique latent distribution for each channel of the image.

Besides, to achieve spatially adaptive deraining for non-uniform rainy images, we propose a spatial density estimation (SDE) module based on a compact dense structure [11]. The proposed SDE module takes a rainy image as input and outputs a density estimation map for each color channel of the rainy image to indicate the rain density on each pixel. The density estimation maps enable adaptive deraining according to the diverse rain densities across spatial locations. This endows the network with the distinguishable capability on different rainy images, sharing similar spirit as the progressive deraining manner [20].

In summary, our major contributions are manifold:

- We propose a variational image deraining (VID) approach under the framework of conditional variational auto-encoder (CVAE) [23], which can learn a mapping from a single input image to many outputs. CVAE effectively perform probabilistic inference and produce diverse predictions. To the best of our knowledge, this is the first work that tackles the deraining problem under the CVAE framework.

- We develop a spatial density estimation (SDE) module based on a dense structure [11]. The SDE module enables our VID method to be adaptive to the rain densities across spatial locations and endows the deraining network with accurate deraining capability.

- We introduce a novel channel-wise strategy for image deraining. The advantages of channel-wise deraining over the whole-channel one are rigorously elaborated under the bright channel prior [33] and demonstrated by extensive ablation studies (please refer to §4.4).

- Extensive experiments on diverse datasets demonstrate that, the proposed CVAE based VID method consistently achieve superior performance to previous state-of-the-art deraining methods, on both synthetic and real-world rainy images. We evaluate the proposed VID method on three diverse synthetic datasets, demonstrating its generalization ability for single image deraining.

2. Related Work

In the past decade, numerous approaches [1, 2, 5, 6, 13, 18, 17, 34, 37, 16, 20, 27] have been proposed to tackle the image deraining problem. Here, we briefly review the related work in this domain.

Traditional methods. For the task of single image deraining, many traditional machine learning methods have been used to solve this problem. Kang et al. [13] decomposed high-frequency parts of rainy images into rainy and non-rainy components. Luo et al. [18] proposed a discriminative sparse coding framework based on image patches. Chen et al. [2] proposed a low-rank appearance model for removing rain streaks. Chang et al. [1] leveraged the low-rank property of rain streaks, and performed deraining by low-rankness based layer decomposition. Li et al. [17] proposed a method that uses simple patch-based priors for both the background and rain layers.

Deep methods. In recent years, many deep learning based deraining methods have achieved promising performance. Fu et al. [5] was the first to introduce deep learning methods (DerainNet) to the deraining problem. They then [6] proposed a deep detail network (DDN) to directly reduce the mapping range from input to output. Yang et al. [34] designed a deep recurrent dilated network (JORDER) to jointly detect and remove rain streaks. Zhang et al. [37] used a generative adversarial network (GAN) to prevent the background image from being degenerated of when extracted from rainy images. They [37] also proposed a density-aware multi-stream densely connected convolutional neural network [11] based algorithm (DID-MDN) for joint rain density estimation and deraining. Li et al. [16] proposed a recurrent squeeze-and-excitation context aggregation net (RESCAN) for the problem of rain streak layers overlapping with each other. Ren et al.[20] present a new baseline network for single image deraining. Wei et al. [27] firstly propose a semi-supervised learning paradigm toward for single image deraining.

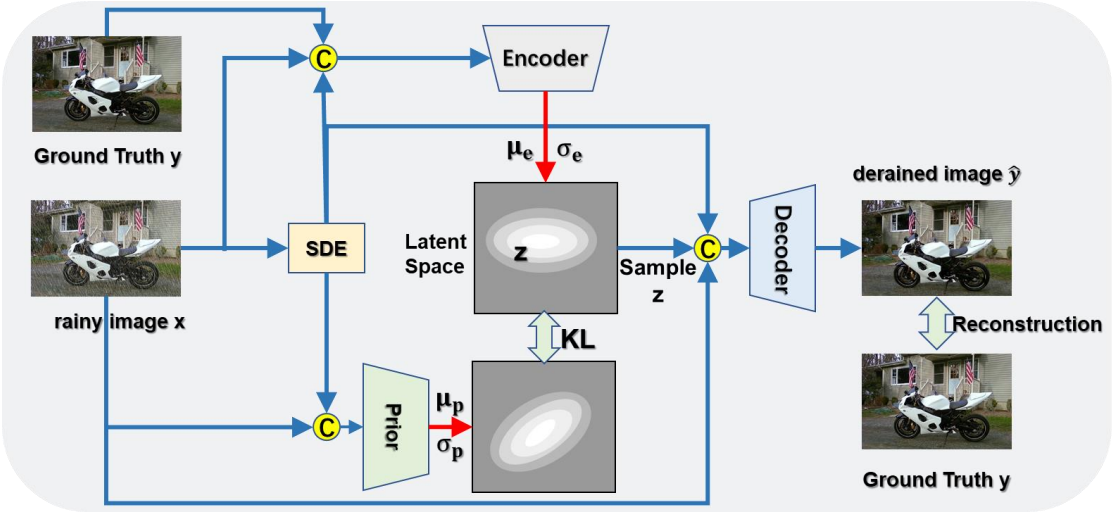


Figure 2: **Illustration of the learning stage of VID.** $\mathbf{x}$ is the conditional rainy image. $\mathbf{y}$ is the clean image. The concatenation of $\mathbf{x}$, $\mathbf{y}$ and density estimation map D_c by the SDE module is the input of the encoder. The decoder produces output $\hat{\mathbf{y}}$ based on D_c and a sampled $\mathbf{z}_e$ from the latent distribution $\mathcal{N}(\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e)$, conditioned also on $\mathbf{x}$. The red arrows indicate the generation of latent space. ⓒ denotes the concatenation.

3. Variational Image Deraining

In this section, we present the proposed Variational Image Deraining (VID) method. In §3.1, we introduce the preliminaries of conditional variational auto-encoder (CVAE). Then we describe the proposed CVAE based image deraining network in §3.2, and present its inference in §3.3. The technical details about the spatial density estimation (SDE) module are provided in §3.4. Finally, we formulate the proposed method via a channel-wise scheme in §3.5.

3.1. Preliminaries on CVAE

Variational auto-encoder (VAE) is a powerful generative framework for learning the latent structure in complex data [15, 21, 10]. The generative process of a VAE is as follows: the encoder takes observable data $\mathbf{x}$ as input and outputs a data-conditional distribution $q(\mathbf{z}|\mathbf{x})$ over a latent vector $\mathbf{z}$. A sample $\mathbf{z} \sim p_\theta$ is drawn from the distribution p_θ and used by the decoder to determine a code-conditional reconstruction distribution $p_\theta(\mathbf{x}|\mathbf{z})$ over the original data $\mathbf{x}$. The objective of VAE is to maximize the variational lower bound of $p_\theta(\mathbf{x})$:

$$\log p_\theta(\mathbf{x}) \geq -D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}), \tag{1}$$

where $\mathbf{z}^{(l)} = g_\phi(\mathbf{x}, \boldsymbol{\epsilon}^{(l)}), \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

As VAE is uncontrolled and unable to generate specific data, its conditional extension (i.e., CVAE) was proposed by Sohn et al. [23] to model latent variables and data, both conditioned on side information such as a part or label of the image. By taking the conditioning $\mathbf{x}$ into account, we can empirically write the lower bound to be maximized as:

$$\begin{aligned} \widetilde{\mathcal{L}}_{\mathrm{CVAE}} = &- \mathrm{KL}(\mathrm{q}_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})||\mathrm{p}_\theta(\mathbf{z}|\mathbf{x})) \\ &+ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x}), \end{aligned} \tag{2}$$

where $\mathbf{z}^{(l)} = g_\phi(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}^{(l)}), \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Here, $p_\theta(\mathbf{z}|\mathbf{x})$ is assumed to be an isotropic Gaussian distribution and $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$, while $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ are Gaussian distributions.

CVAE has shown its great power in diverse computer vision tasks, such as trajectory prediction [25], image colorization [3], image generation [4], and multi-modal human dynamic generation [32]. In this work, to the best of our knowledge, we are among the first to explore the potential power of the generative CVAE model for low-level vision tasks such as image deraining.

3.2. Learning CVAE for Image Deraining

Single image deraining is essentially an ill-posed problem and highly non-trivial for generating optimal solutions in realistic rainy images. VAE has the innate capability of modeling latent distributions, which can be used to find the distributions of clean images. However, it cannot be directly applied for deraining, since it can only take in and output the same rainy image, while cannot output a derained image from the rainy input. Since CVAE is able to generate specific output (e.g., derained image) from the input (e.g., rainy image), it is employed for image deraining in this work.

As shown in Fig. 2, the proposed CVAE based deraining framework is consisted of an encoder, a prior network, and a decoder. Conditioned on the rainy image $\mathbf{x}$, the encoder learns to map its corresponding clean image $\mathbf{y}$ into a latent distribution $\mathcal{N}(\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e)$ that carries information about the clean image distribution. To guarantee that the sam-

Algorithm 1 Variational Image Deraining (VID)

Learning: Input pairs of rainy and clean images $\{\mathbf{x}, \mathbf{y}\}_{i=1}^{N}$
θ_s, ϕ, θ_p, $\theta_d \leftarrow$ Initialize parameters
repeat
SDE: $\hat{D}_c \leftarrow \mathrm{SDE}_{\theta_s}(\mathbf{x})$
Encoder: $\begin{cases} \boldsymbol{\mu_e}, \boldsymbol{\sigma_e} \leftarrow E_\phi(\mathbf{x}, \mathbf{y}, \hat{D}_c) \\ \mathbf{z}_e \leftarrow \boldsymbol{\mu_e}(\mathbf{x}) + \boldsymbol{\epsilon} * \boldsymbol{\sigma_e}(\mathbf{x}), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{cases}$
Prior: $\begin{cases} \boldsymbol{\mu_p}, \boldsymbol{\sigma_p} \leftarrow P_{\theta_p}(\mathbf{x}, \hat{D}_c) \\ \mathbf{z}_p \leftarrow \boldsymbol{\mu_p}(\mathbf{x}) + \boldsymbol{\epsilon} * \boldsymbol{\sigma_p}(\mathbf{x}), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{cases}$
Decoder: $\hat{\mathbf{y}} \leftarrow D_{\theta_d}(\mathbf{x}, \mathbf{z}, \hat{D}_c)$
$\mathbf{g} \leftarrow \nabla_{\theta_s, \phi, \theta_p, \theta_d} \mathcal{L}(\theta_s, \phi, \theta_p, \theta_d; \mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon})$
θ_s, ϕ, θ_p, $\theta_d \leftarrow$ Update parameters using gradients $\mathbf{g}$
until convergence
return θ_s, ϕ, θ_p, θ_d

Inference: Input rainy image $\mathbf{x}$
SDE: $\hat{D}_c \leftarrow \mathrm{SDE}_{\theta_s}(\mathbf{x})$
Prior: $\begin{cases} \boldsymbol{\mu_p}, \boldsymbol{\sigma_p} \leftarrow P_{\theta_p}(\mathbf{x}, \hat{D}_c) \\ \mathbf{z}_p \leftarrow \boldsymbol{\mu_p}(\mathbf{x}) + \boldsymbol{\epsilon} * \boldsymbol{\sigma_p}(\mathbf{x}), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{cases}$
Decoder: $\hat{\mathbf{y}} \leftarrow \frac{1}{S} \sum_{s=1}^{S} D_{\theta_d}(\mathbf{x}, \mathbf{z}, \hat{D}_c)$
return Derained image $\hat{\mathbf{y}}$

pled latent variable $\mathbf{z}$ from the latent distribution is related to the input $\mathbf{x}$ during inference, we introduce a prior network to make sure that the latent distribution obtained by learning is consistent with that obtained by inference. The prior network learns to map a rainy image $\mathbf{x}$ into another latent distribution $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p)$ that posses information about the rainy image distribution. The decoder then reconstructs the derained image $\hat{\mathbf{y}}$ based on a sampled $\mathbf{z}_e$ from the latent distribution $\mathcal{N}(\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e)$, conditioned also on the rainy image $\mathbf{x}$. To compute the gradient more amenably, we use reparameterization technique [15] to sample $\mathbf{z}$, i.e., $\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\epsilon} * \boldsymbol{\sigma}(\mathbf{x})$, where $\boldsymbol{\epsilon}$ is sampled from an auxiliary noise distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

We need to maximize the conditional variational lower bound defined in (2) for learning. The first term in (2) acts as a regularization term that pushes $q_\phi(\mathbf{z}_e|\mathbf{x}, \mathbf{y})$ to match the prior distribution $p_\theta(\mathbf{z}_p|\mathbf{x})$. We take Kullback-Leibler (KL) divergence as the penalty function to minimize the gap between the Gaussian distributions $q_\phi(\mathbf{z_e}|\mathbf{x}, \mathbf{y})$ and $p_\theta(\mathbf{z_p}|\mathbf{x})$. The second term in (2) is the reconstruction error, which measures the information loss between the sampled latent code $\mathbf{z}_e$ and the clean image. We maximize the conditional log-likelihood $\mathbb{E}_{q_\phi(\mathbf{z_e}|\mathbf{x},\mathbf{y})}[\log p_\theta(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{z_e})]$ for accurate reconstruction. In practice, the reconstruction error can be computed as the L_2 loss between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

3.3. Inference

To obtain a deterministic output during inference, we draw S latent codes $\{\mathbf{z}_p^{(s)}\}_{s=1}^{S}$ from the prior distribution $p_\theta(\mathbf{z}_p|\mathbf{x})$ using the prior network, and take the average of the posteriors to make a prediction. We compute the marginal likelihood using the Monte Carlo method [22]:

$$p_\theta(\mathbf{y}|\mathbf{x}) \approx \frac{1}{S} \sum_{s=1}^{S} p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{z}_p^{(s)}), \mathbf{z}_p^{(s)} \sim p_\theta(\mathbf{z}_p|\mathbf{x}). \quad (3)$$

We use the Monte Carlo sampling to estimate the conditional likelihoods (CLL) of the second term of (2). We find that 100 samples are enough to obtain an accurate estimate of the CLL in our experiments (please refer to (3)). We summarize the learning and inference procedures of the proposed VID method in Algorithm 1.

Loss for CVAE. The CVAE is trained to maximize the conditional log-likelihood of the second term of (2). Since this objective function is intractable, we instead maximize the variational lower bound in (2). We minimize the KL divergence between the distribution $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ and the prior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ to mitigate the discrepancies between the encoding of latent variables at learning and inference stage as follows:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{i=1}^{N} q_\phi(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i) \log\left(\frac{q_\phi(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)}{p_\theta(\mathbf{z}_i|\mathbf{x}_i)}\right), \quad (4)$$

where $q_\phi(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)$=$\mathcal{N}(\boldsymbol{\mu}_e, \boldsymbol{\sigma}_e)$, $p_\theta(\mathbf{z}_i|\mathbf{x}i)$=$\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p)$.

To maximize $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{z})]$ for the reconstruction, we define the loss $\mathcal{L}_{\mathrm{rec}}$ as the ℓ_2 loss between clean image $\mathbf{y}$ and derained image $\hat{\mathbf{y}}$ as follows:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c \in \{r,g,b\}} ||\mathbf{y}_{i,c} - \hat{\mathbf{y}}_{i,c}||_F^2, \quad (5)$$

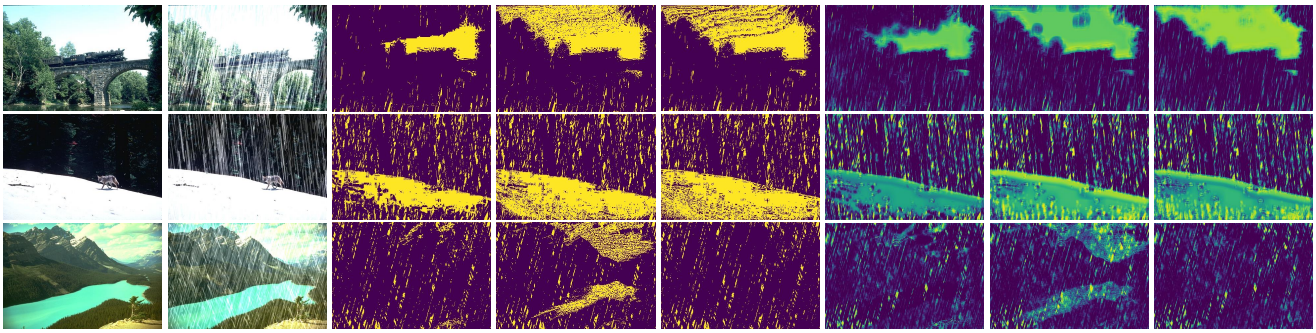
where $\hat{\mathbf{y}}_{i,c} = f_c^{\mathrm{rec}}(\mathbf{x}_{i,c}, \mathbf{y}_{i,c}, D_{i,c})$ is the CVAE associated with the c-th channel. The CVAE takes each individual color channel of the rainy image $\mathbf{x}$, clean image $\mathbf{y}$ and the rain density estimation map D_c in channel c as the inputs, and outputs the derained image $\hat{\mathbf{y}}_c$ of that channel. Taking the $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_{\mathrm{rec}}$ together, we obtain the overall loss $\mathcal{L}_{\mathrm{CVAE}}$ as follows:

$$\mathcal{L}_{\mathrm{CVAE}} = \mathcal{L}_{\mathrm{rec}} + \beta \mathcal{L}_{\mathrm{KL}}, \quad (6)$$

where $\beta > 0$ is a regularization parameter.

3.4. Spatial Density Estimation

The rain streaks are usually unevenly distributed on a rainy image, vary across different spatial locations. The methods ignoring the spatial variance will inevitably generate inaccurate deraining results on the unevenly distributed rainy images. Although global density information is considered in [36] by grading rain strength into different levels, inaccurate deraining results are still unavoidable in local regions. Specifically, since the rain streaks are usually randomly distributed in the rainy image, it is difficult to locate the rainy regions consistently.



(a) Ground Truth (b) Rainy image (c) D_r (d) D_g (e) D_b (f) $\hat{D}_r$ (g) $\hat{D}_g$ (h) $\hat{D}_b$

Figure 3: **Illustration of rain distributions and rain density estimation maps on the r, g, b channels, respectively.** (a) and (b) are rainy and corresponding clean images, respectively. (c), (d) and (e) show rain distributions on r, g, b channels, respectively. (f), (g) and (h) are the generated density estimation maps for the three color channels, respectively.

We tackle the challenge by proposing a spatial density estimation (SDE) module, and embed it into the proposed VID method to make it spatially adaptive for deraining. By this way, the pixels with strong rain streaks will be restored with more emphasis, while those with weak rain streaks will be slightly restored. Specifically, the proposed SDE module is designed as a compact densely-connected convolutional block with five layers [11] to learn a density estimation map for the input rainy image, indicating the density distribution of the rain streaks at different spatial locations. It takes the whole rainy image as input and outputs a density estimation map specific for each color channel.

The learning of the density estimation maps is performed in a fully supervised manner. Specifically, we subtract a rainy image $\mathbf{x}$ from its corresponding clean image $\mathbf{y}$ (taken as "ground truth"), and produce a residual image denoted as $\hat{R}$. $\hat{R}_c$ indicates a color channel in R, where $c \in \{r, g, b\}$, and $\hat{R}_c(x)$ denotes a pixel value of position x on each channel. $\hat{R}_c(x) = 0$ indicates that there is no rain at this pixel, while $\hat{R}_c(x) \neq 0$ indicates there is rain at this pixel. Based on the residual map $\hat{R}$, we generate the ground truth image for supervised learning of density estimation maps using:

$$D_c(x) = \begin{cases} 0 & \hat{R}_c(x) \neq 0 \\ 1 & \hat{R}_c(x) = 0 \end{cases}, \tag{7}$$

where D_c is the ground truth for the c-th channel.

We plot D_c in Fig. 3 (c), (d), and (e) for the three color channels r, g and b, respectively. As can be seen, the rain streaks are distributed randomly across spatial locations and the distribution patterns on three channels are very different. The main reason is that, the light emitted from different sources are in very different strength. Some examples are shown in Fig. 3 can be the sunlight (first row), the white floors (second row), and the green water (third row). We also plot the density estimation maps learned by the proposed SDE module in Fig. 3 (f), (g), and (h), for the r, g, and b color channel, respectively. As can be seen, the density estimation maps learned by the proposed SDE module are close to those of the ground truth, which indicates that the SDE module can accurately localize the rain regions for the three color channels. Therefore, the proposed CVAE based VID method can obtain reasonable density estimation maps for adaptive deraining effectiveness according to the rain strength in different regions.

Loss for SDE. The SDE module is trained in a supervised way. It takes the whole color image $\mathbf{x}$ as input and generates the density estimation maps D_c for each color channel. The loss function of the SDE module takes the following form:

$$\mathcal{L}_{\text{SDE}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c \in \{r,g,b\}} ||D_{i,c} - \hat{D}_{i,c}||_F^2, \tag{8}$$

where $\hat{D}_{i,c} = f_c^D(\mathbf{x}_{i,c})$ and $f_c^D(\cdot)$ is the SDE module associated with the c-th channel. This loss minimizes the difference between the generated density estimation maps and ground truth. The obtained density estimation maps are passed to the CVAE as the inputs, which steers the CVAE to focus more on regions with rain streaks.

3.5. Channel-wise Deraining

The bright pixels in rainy images tend to be duller if we treat the R, G, B channels equally. This is due to that the rain density are in different distributions for different color channels, which has not be explored in previous deraining methods. With the bright channel prior (BCP) [33], we demonstrate that the channel-wise deraining scheme can help the proposed VID better preserve the pixel brightness in derained images than previous methods which treat different channels equally.

BCP prior [33] describes an observation that in most natural scenes, at least one color channels possesses high intensities for each pixel. The BCP prior is defined as

$$J^{bright}(x) = \max_{y \in \Omega(x)} (\max_{c \in \{r,g,b\}} J^c(y)), \tag{9}$$

where J^c is a color channel of image J and $\Omega(x)$ is a local patch centered at location x. The intensity of J^{bright} should

be close to 1 (intensity is in $[0, 1]$), except in a situation lacking light or dominated by shadow [33]. Based on the BCP, we propose a proposition to theoretically validate the proposed channel-wise deraining strategy as follows:

Proposition 1. *Denote the $\bar{B}$ and B as images derained without and with distinguishing different color channels, respectively. Then, the intensity of the pixels in $\bar{B}$ is much lower than B. That is, the number of brightest pixels in $\bar{B}$ tends to be less than that in B. To be more precise, we have*

$$||1 - B||_0 < ||1 - \bar{B}||_0. \tag{10}$$

The proof of this proposition is provided in the *Supplementary Files.*

3.6. Optimization

The proposed VID model (3) is optimized by jointly minimizing the negative conditional variational lower bound (2) and the loss of the SDE module (3.4). Specifically, we formulate the objective function (11) as a multi-task optimization problem.

Taking the losses in (6) and (8) together, we obtain the overall multi-task loss as follows:

$$\mathcal{L}_{\text{VID}} = \mathcal{L}_{\text{CVAE}} + \lambda \mathcal{L}_{\text{SDE}}, \tag{11}$$

where $\lambda > 0$ is a regularization parameter to balance the importance of $\mathcal{L}_{\text{SDE}}$ and $\mathcal{L}_{\text{CVAE}}$. However, we observed that we constantly obtain peak performance when we treat them equally, i.e., $\lambda = 1$. In our VID, the SDE module and CVAE are jointly trained by gradient decent via backward error propagation, in an end-to-end framework.

4. Experiments

In this section, we conduct extensive experiments to demonstrate the performance of the proposed variational image deraining (VID) method. We also conduct comprehensive ablation studies to study its effectiveness.

4.1. Experimental Protocol

Implementation Details In the learning stage, we randomly generate 2,000 pairs of image patches with size 64×64, from each training set. For the SDE module, we set the filter size as $s_1 = 3$, and the number of filters as $a_1 = 16$. Each convolutional layer is followed by the batch normalization [12], and the ReLU [19] activation operations. For the last layer, we use the sigmoid activation function to make the density estimation map within the range of $[0, 1]$. In the CVAE, we set the filter size as $s_2 = 3$, and the number of convolution filters as $a_2 = 16$ in the encoder and the prior network. In the last layer of the encoder and prior network, the first half is $\boldsymbol{\mu}$ and another half is $\boldsymbol{\sigma}$. We set the number of convolution filters as $a_3 = 1$. For the decoder, we set the filter size as $s_3 = 3$, and the number of deconvolution filters as $a_4 = 16$. We set the depth as $L = 7$ for encoder, decoder and prior networks, and employ Leaky ReLU [29] as activation function. Each layer is also followed by a batch normalization layer in the CVAE. We use the Adam optimizer [14] with default parameters, at a weight decay of 10^{-10} and a mini-batch size of 32. The initial learning rate is 0.01, and divided by 10 at each epoch.
Comparison Methods. We compare the proposed VID method with 5 state-of-the-art deraining methods, including Deep Detail Network (DDN) [6], Joint Rain Detection and Removal (JORDER) [34], Density-aware Deraining (DID-MDN) [36], and Recurrent Squeeze-and-excitation Context Aggregation Network (RESCAN) [16].
Evaluation Metrics. Following previous works [6, 34, 36, 16], we adopt two commonly-used metrics, i.e., peak signal to noise ratio (PSNR) and structure similarity index (SSIM) [26], to measure the performance of deraining on the synthesized datasets. Since the real-world rainy images have no "ground truth" images, we can only show the comparisons on the visual quality of derained images by different image deraining algorithms.

4.2. Results on Synthetic Rain Removal

Datasets. We perform experiments on 3 synthetic datasets and 1 real-world dataset. The first dataset is provided in [6] and contains 14,000 synthesized clean/rainy image pairs. Following the settings in [36], 13,000 images are used for learning, and the remaining 1000 images are used for testing (denoted as *T1*). The second synthesized dataset is provided in [34] and consists of 1,800 pairs of heavy rain images and 200 pairs of light rain images for learning. The two sets (*Rain100L* and *Rain100H*) are used for testing (denoted as *T2*). The third dataset [36] contains 12,000 synthesized clean/rainy image pairs, which includes 4,000 heavy rainy images, 4,000 medium rainy images, 4,000 light rainy images. The 1,200 pairs of clean/rainy images for testing are denoted as *T3*. As far as we know, this is the first work that conducts experimental evaluation on all these three datasets.
SSIM and PSNR. The quantitative comparisons are reported in Table 1. Our VID method substantially exceeds previous methods on all three datasets. In particularly, on *T1*, our method outperforms the second best method by 4% and 3.6 in terms of SSIM and PSNR, respectively. The superior performance demonstrates the great effectiveness of our method for single image deraining.
Visual Quality. In Fig. 4, we show the comparison of the visual quality of different methods. It can be seen that, our VID method produces removes rain streaks more clearly, and preserve better image details than previous methods.

4.3. Results on Realistic Rain Removal

We also apply the proposed VID on removing the rain streaks in real rainy photographs. The proposed CVAE

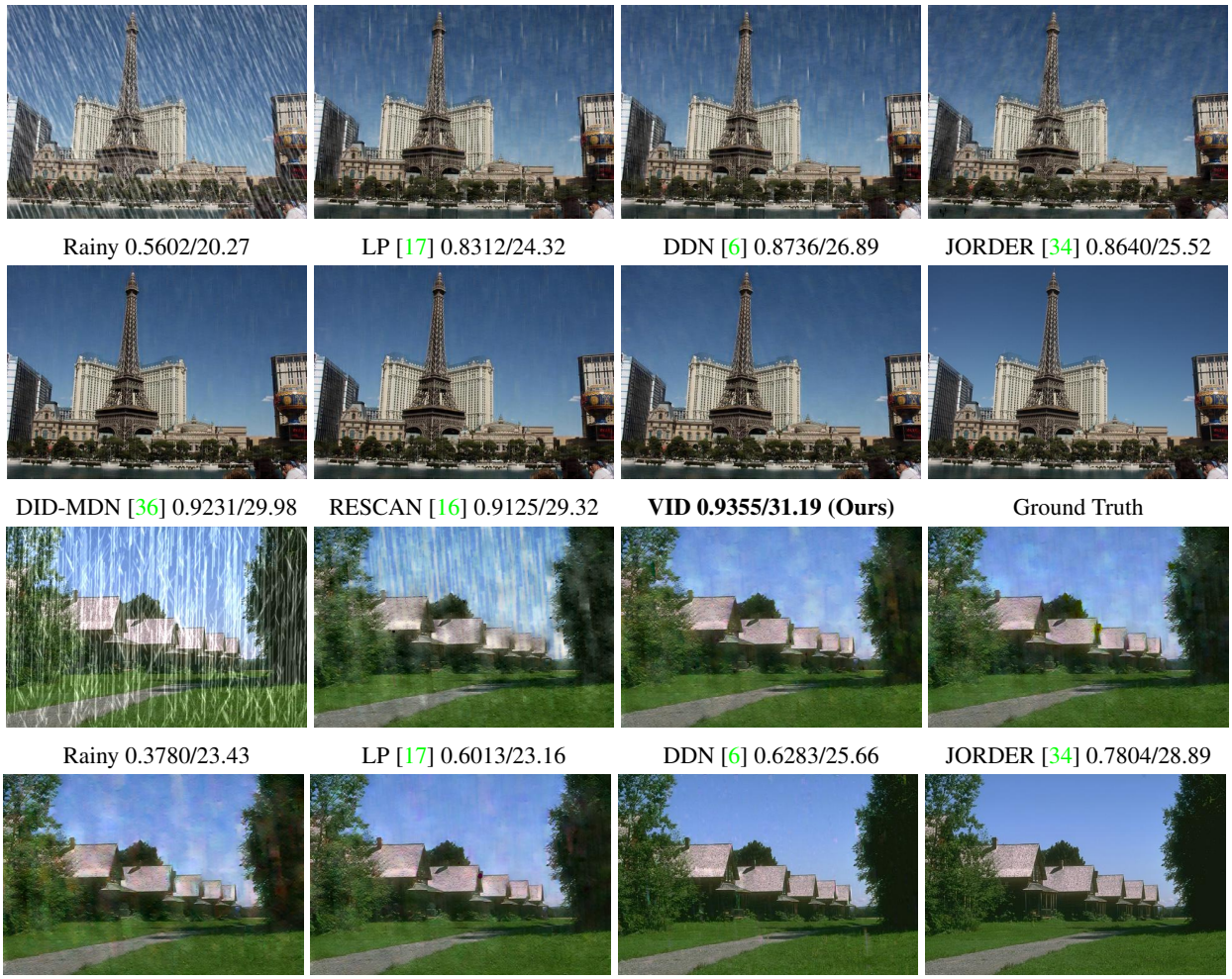


Figure 4: **Comparisons of derained images by different methods on synthetic datasets [6] and [34].**

Dataset		Input	LP [17]	DDN [6]	JORDER [34]	DID-MDN [36]	RESCAN [16]	VID (Ours)
T1		0.7695/19.31	0.8312/24.35	0.8851/25.63	0.8405/22.36	0.9092/26.07	0.8812/25.45	**0.9325/28.73**
T2	*L*	0.8332/23.52	0.8253/24.14	0.8494/25.84	0.8835/28.32	0.8725/27.13	0.9058/29.17	**0.9343/32.10**
	H	0.3702/12.13	0.5444/14.26	0.6928/22.26	0.7382/23.45	0.7315/23.25	0.8458/26.45	**0.8721/27.89**
T3		0.7781/21.15	0.8514/25.23	0.8978/27.33	0.8622/24.32	0.9087/27.95	0.8712/26.19	**0.9326/30.82**

Table 1: **Quantitative comparison of different methods** in terms of SSIM and PSNR(dB) on synthesized test datasets D1 [6], D2 [34], and D3 [36]. The terms "*L*" and "*H*" denotes the "Rain100L" and "Rain100H", respectively.

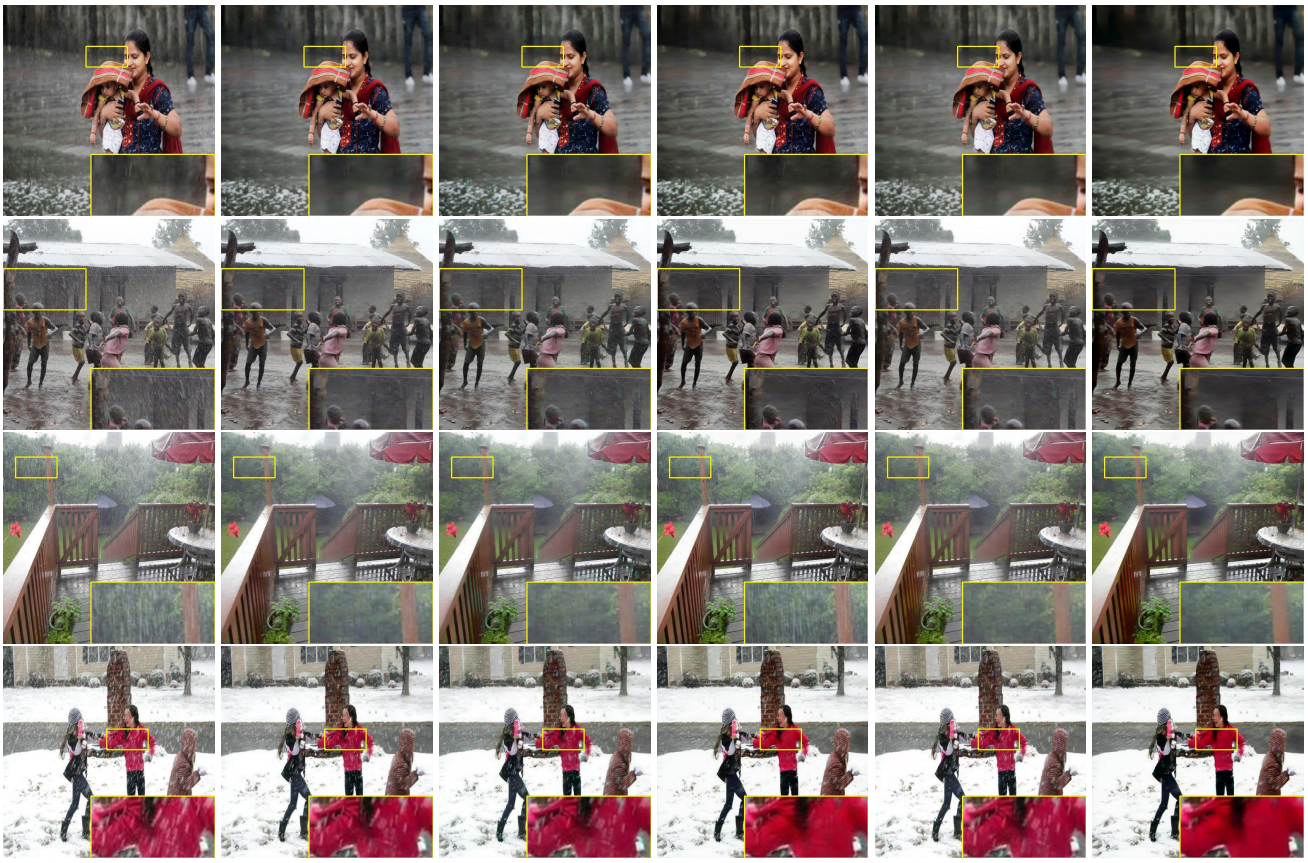
based VID model is learned on the dataset used in DID-MDN [36]. We use the rainy images in [35], including 4 different representative scenarios (shown in Figs. 5 (a)): light rain, medium rain, heavy rain, and snow (from top to bottom). As shown in Figs. 5 (b)-(f), the proposed VID outperforms all previous methods on real rainy images. More results can be found in the *Supplemental File*.

4.4. Ablation Study

We conduct more detailed ablation studies of our proposed VID on image deraining. We assess the effectiveness of 1) the CVAE model, 2) the SDE module, and 3) the channel-wise scheme. DDN [6] is employed as baseline.

1) CVAE model. As shown in Table 2, the performance of channel-wise CVAE is 0.9154/30.14dB on SSIM/PSNR, much better than that of channel-wise DDN (0.8763/28.19dB), demonstrating the effectiveness of the proposed CVAE based method for image deraining.

2) SDE module. We can see in Table 2 that both Channel-wise DDN with SDE and VID perform much better than channel-wise DDN and channel-wise CVAE (without SDE), which verifies the effectiveness of the SDE module.



(a) Rainy Image (b) DDN [6] (c) JORDER [34] (d) DID-MDN [36] (e) RESCAN [16] (f) **VID (Ours)**

Figure 5: **Comparisons of derained images by different methods on real-world rainy images** from the datasets [35]. We choose four types of rainy image, which include representative scenarios: light rain, medium rain, heavy rain, and snow respectively from top to bottom.

	DDN [6]	Channel-wise DDN	Channel-wise DDN + SDE	Channel-wise CVAE	**VID (Ours)**
Channel-wise	✘	✓	✓	✓	✓
SDE	✘	✘	✓	✘	✓
CVAE	✘	✘	✘	✓	✓
Rain100L	0.8494/25.84	0.8763/28.19	0.8913/29.23	0.9154/30.14	**0.9343/32.10**
Rain100H	0.7270/22.95	0.7637/24.39	0.7817/26.11	0.8332/26.35	**0.8721/27.89**

Table 2: **Quantitative comparisons of different variants** in terms of SSIM and PSNR (dB) on dataset *T2* [34].

3) Channel-wise deraining. From Table 2, it is easy to see that the baseline DDN is dramatically improved by conducting channel-wise deraining. This clearly demonstrates the advantage of channel-wise deraining over methods that do not distinguish color channels during deraining.

5. Conclusion

In this paper, we proposed to tackle image deraining in the conditional variational auto-encoder (CVAE) framework. CVAE models the latent distributions of image priors, from which the clean images are generated for image deraining. Moreover, we introduced a channel-wise scheme to achieve the image deraining more adaptive in different color channels. A spatial density estimation module is developed to achieve spatially adaptive deraining performance on uneven rainy images. Experiments on both synthetic and real-world datasets show that our method achieves superior performance to previous state-of-the-art deraining methods.

Acknowledgment

This work was supported by the National Natural Science Foundation of China(61871016, 61976060), Project of Educational Commission of Guangdong province of China(2018KCXTD019).

References

[1] Y. Chang, L. Yan, and S. Zhong. Transformed low-rank model for line pattern noise removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1726–1734, 2017. 2

[2] Y.-L. Chen and C.-T. Hsu. A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1968–1975, 2013. 2

[3] A. Deshpande, J. Lu, M.-C. Yeh, M. Jin Chong, and D. Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6837–6845, 2017. 3

[4] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 3

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

[6] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley. Removing rain from single images via a deep detail network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1715–1723, 2017. 1, 2, 6, 7, 8

[7] X. Fu, D. Zeng, Y. Huang, X. Ding, and X.-P. Zhang. A variational framework for single low light image enhancement using bright channel prior. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 1085–1088. IEEE, 2013. 2

[8] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 1

[9] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2011. 2

[10] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013. 3

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017. 2, 5

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 6

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

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

[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 4

[16] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *European Conference on Computer Vision*, pages 262–277. Springer, 2018. 1, 2, 6, 7, 8

[17] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2736–2744, 2016. 2, 7

[18] Y. Luo, Y. Xu, and H. Ji. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3397–3405, 2015. 2

[19] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. 6

[20] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng. Progressive image deraining networks: A better and simpler baseline. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[21] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 3

[22] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method*. Wiley, Dec. 2007. 1, 2, 4

[23] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015. 1, 2, 3

[24] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1

[25] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016. 3

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[27] W. Wei, D. Meng, Q. Zhao, Z. Xu, and Y. Wu. Semi-supervised transfer learning for image rain removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[28] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 1

[29] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 6

[30] J. Xu, L. Zhang, and D. Zhang. A trilateral weighted sparse coding scheme for real-world image denoising. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[31] J. Xu, L. Zhang, D. Zhang, and X. Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1096–1104, 2017. 2

[32] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, and H. Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018. 3

[33] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao. Image deblurring via extreme channels prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 6, 2017. 2, 5, 6

[34] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2017. 1, 2, 6, 7, 8

[35] H. Zhang and V. M. Patel. Convolutional sparse and low-rank coding-based rain streak removal. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 1259–1267. IEEE, 2017. 7, 8

[36] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. *arXiv preprint arXiv:1802.07412*, 2018. 1, 4, 6, 7, 8

[37] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017. 2